2m11.3168.6

Université de Montréal

Routing and dimensioning of 3G multi-service networks

par

Raha Pooyania

11507737

Département d'informatique et
de recherche opérationnelle

Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de
Maître ès sciences (M. Sc.)
en informatique

Avril, 2004

Copyright © Raha Pooyania, 2004

Faculté des Études Supérieures
Grade octroyé
à compter du
03 JUIN 2004
Université de Montréal

QA
76
U54
2004
v.024

Université de Montréal

Direction des bibliothèques

AVIS

L'auteur a autorisé l'Université de Montréal à reproduire et diffuser, en totalité ou en partie, par quelque moyen que ce soit et sur quelque support que ce soit, et exclusivement à des fins non lucratives d'enseignement et de recherche, des copies de ce mémoire ou de cette thèse.

L'auteur et les coauteurs le cas échéant conservent la propriété du droit d'auteur et des droits moraux qui protègent ce document. Ni la thèse ou le mémoire, ni des extraits substantiels de ce document, ne doivent être imprimés ou autrement reproduits sans l'autorisation de l'auteur.

Afin de se conformer à la Loi canadienne sur la protection des renseignements personnels, quelques formulaires secondaires, coordonnées ou signatures intégrées au texte ont pu être enlevés de ce document. Bien que cela ait pu affecter la pagination, il n'y a aucun contenu manquant.

NOTICE

The author of this thesis or dissertation has granted a nonexclusive license allowing Université de Montréal to reproduce and publish the document, in part or in whole, and in any format, solely for noncommercial educational and research purposes.

The author and co-authors if applicable retain copyright ownership and moral rights in this document. Neither the whole thesis or dissertation, nor substantial extracts from it, may be printed or otherwise reproduced without the author's permission.

In compliance with the Canadian Privacy Act some supporting forms, contact information or signatures may have been removed from the document. While this may affect the document page count, it does not represent any loss of content from the document.

Université de Montréal

Faculté des études supérieures

Ce mémoire intitulé :

Routing and dimensioning of 3G multi-service networks

présenté par :

Raha Pooyania

a été évalué par un jury composé des personnes suivantes :

Michel Gendreau

(président-rapporteur)

Brigitte Jaumard

(directeur de maîtrise)

Bernard Gendron

(membre du jury)

Mémoire accepté le

28 avril 2004

Sommaire

La demande élevée de bande passante pour accéder aux différents types d'application, en particulier les applications multimédia, à partir d'une station mobile tout en satisfaisant une certaine qualité de service correspond au nouveau visage des réseaux sans fil de troisième génération. Le besoin accru en ressources de la part des applications multimédia, attire l'attention sur les outils de dimensionnement.

Le but de ce mémoire de M.Sc. est la généralisation d'un modèle mathématique développé dans la thèse de M.Sc. de C. Voisin pour le dimensionnement de réseaux 3G en ce qui concerne les aspects de routage. En effet, l'objectif est d'étudier l'impact d'une politique de multi-routage versus une politique de mono-routage sur le coût de dimensionnement.

Le modèle généralisé correspond à un programme mathématique linéaire mixte en variables 0-1. La validation et les expériences informatiques comparatives ont été faites sur divers exemples de trafic avec différentes caractéristiques. Bien que des expériences aient été faites seulement sur des réseaux de taille limitée, il est déjà possible de mettre en évidence un avantage clair du multi-routage en comparaison du mono-routage.

Mots clés : Réseau 3G, CDMA2000, Dimensionnement, Multi-routage, Contrôle d'admission, Soft handoff, Capacité radio, GdS, QdS.

Abstract

The high demand of bandwidth for access to different kinds of applications, especially multimedia applications, by a mobile station, while satisfying their requested Quality of Service, is the new face of third generation of networks. The need of multimedia applications for high bandwidth, convey the researches to dimensioning tools.

The purpose of this M.Sc. Thesis is the generalization of a mathematical model developed in the M.Sc. Thesis of C. Voisin for the dimensioning of 3G networks with respect to the routing aspects. Indeed, the objective is to study the impact of a multi versus a mono routing path policy on the dimensioning cost.

The generalized model corresponds to a linear mixed mathematical program with 0-1 variables. Validation and comparative computational experiences have been performed on various traffic instances with different characteristics. Although experiments have been performed on a network with limited size, there is already a clear advantage of multi-routing over mono-routing.

Keywords : 3G Network, CDMA2000, Dimensioning, Multi-routing, Call Admission Control, Soft handoff, Radio capacity, GoS, QoS.

Table of contents

List of tables

List of figures

List of abbreviations

2G	Second Generation Cellular Mobile Systems
3G	Third Generation Cellular Mobile Systems
4G	Fourth Generation Cellular Mobile Systems
AAA	Authentication, Authorization and Accounting
AMPS	Advanced Mobile Phone System
BGP	Border Gateway Protocol
BS	Base Station
BSC	Base Station Controller
CAC	Call Admission Control
CBR	Constant Bit Rate
CDMA	Code Division Multiple Access
CN	Core Network
DL	DownLink
DS-WCDMA	Direct Sequence Wide-band CDMA
FA	Foreign Agent
FCFS	First Come First Served
FDD	Frequency Division Duplex
FDMA	Frequency Division Multiple Access
FER	Frame Error Rate
FIFO	First In First Out
FTP	File Transfer Protocol
GoS	Grade of Service

GPRS	General Packet Radio Service
GPS	Global Positioning System
GPS	General Processor Sharing
GSM	Global System for Mobile Communication
HA	Home Agent
INTSERV	Integrated Services
IP	Internet Protocol
MC-CDMA	Multi Carrier CDMA
MIP	Mobile Internet Protocol
MS	Mobile Station
MSC	Mobile Switching Center
OSPF	Open Shortest Path First
PCF	Packet Control Function
PDSN	Packet Data Serving Node
PGPS	Packet by packet General Processor Sharing
PPP	Point to Point Protocol
PSTN	Public Switched Telephone Network
QoS	Quality of Service
RAB	Radio Access Bearer
RAN	Radio Access Network
RIP	Routing Information Protocol
RN	Radio Network
RPPS	Rate Proportional Processor Sharing
RRC	Radio Resource Control
RTP	Real-time Transport Protocol
SIR	Signal to Interference Ratio
TCP	Transmission Control Protocol
TDD	Time Division Duplex
TDMA	Time Division Multiplex Access
UDP	User Datagram Protocol

UL	UpLink
UMTS	Universal Mobile Telecommunications System
VBR	Variable Bit Rate
WCDMA	Wide-band Code Division Multiple Access
WFQ	Weighted Fair Queuing
WWW	World Wide Web

In the name of the Lord of wisdom and mind,
To nothing sublimer can thought be applied.
Ferdosi 940 A.D.

Acknowledgments

My sincere gratitude to my supervisor Professor Brigitte Jaumard for her enthusiasm, integral view on research, stimulation suggestion, rich comments, kindness, understanding and the financial support.

I am really glad that I had this opportunity to know my friend, Dr. Christophe Meyer, who helped me all the way in this thesis. Thanks for valuable discussion, constant encouragement and availability.

My appreciation to the jury committee, Professor Michel Gendreau and Professor Bernard Gendron who accepted to read my thesis and for their useful comments which helped me to complete this study.

Thanks to the Département d'informatique et recherche opérationnelle (IRO) and the Centre de recherche sur les transports (CRT) for their support and the excellent environment for studying.

A special thank to my great parents, who taught me the good things that really matter in life. I deeply appreciate your unconditional love, support, sacrifices and being a source of pride for me. I thank my amazing sister, my kind brother and brother-in-law and my lovely nephew for their pure love and encouragement.

To my caring grand-parents, memory of my late grand-parents and other members of family and friends for their inspiration and moral support. Thanks to all my friends in Iran and Canada, Ali, Akbar, Hichem, Maryam, Mahboobeh, Nourchen, Rose, Shiva, Salim... for their precious friendship.

And at last to my beloved husband, Vahid, whose endless support has allowed me to chase a dream. Thanks for your love, patience, help, understanding and believing in me through all this long process.

Chapter 1

Introduction

1.1 Motivation

Communication has always been an essential part of every kind of human society. So far, many technologies have been developed for this purpose. Among these, mobile communication has been one of the most important technologies that man has ever used. The abilities of the second generation telecommunication systems, such as GSM (Global System for Mobile Communication), are limited to digital wireless voice traffic. These systems have been designed for voice communications with low-bit-rate data services. Any enhancement or addition of new services also affects the service. Growing demands for transferring high quality images, video and wireless Internet access with high data rate (up to 2 mbps) and needs for data-rich, multimedia services accessed instantly over mobile handsets forced the technology to move to Third Generation Telecommunication Systems (3G) and Fourth Generation Telecommunication Systems (4G).

Every telecommunication operator, developer or vendor in the world is affected by this technology since telecommunications evolve toward a new generation of networks, services and applications. The third and fourth generations of networks are the new faces of wireless network technologies, which have been significantly improved in terms of system capacity, voice quality, and ease of use.

In Japan, telecommunication systems are very close to 4G, based on WCDMA, whose main advantage over 3G is the availability of higher bandwidth. While in North America, the technology is still with 3G. Among the 3G standards, WCDMA (Wide-band Code Division Multiple Access) and CDMA2000 are the most used third generation air interfaces. 3G combines high-speed mobile access with Internet Protocol (IP) based services but it does not concern a super fast connection of mobile communications.

3G is expected to support enhanced multi-media such as voice, data, video and remote control in all known modes such as cellular telephone, e-mail, fax, videoconferencing, web browsing and other services.

The demand of bandwidth is obvious for all kinds of applications. The fixed network can handle the high data rate, but not the Quality of Service (QoS) demands for the new multi-media applications. The goal of the first applications over Internet, such as FTP (File Transfer Protocol), was to have a reliable connection without considering the delay. The best effort connection cannot satisfy the QoS demands of each application either. The new real-time multi-media applications such as videophone, not only need a high data rate but are also very sensitive to delay. Therefore, we need mechanisms to provide the requested Quality of Service for each type of applications. The need for a high data rate and QoS in radio links is another challenge for 3G networks. This is why in 3G networks the Code Division Multiple Access (CDMA) technology which diversifies the bandwidth on the radio links has been used.

In order to satisfy multi-media applications demands, more base stations and advanced equipments are needed. Considering the high prices of these equipments, the use of dimensioning tools with multi-routing strategies are necessary. Therefore, developing mathematical models for network dimensioning that minimize the cost of link capacities in the core network as well as the equipments in the radio network under different routing strategies are the objectives of this project. For each requested session from the mobile station, there may be several potential base stations which are ready to support this session and are connected to different Base Station Controllers (BSC). In addition it is assumed that there may be different routing paths between a given pair of origin and destination. This means multi-routing, which is a complex subject with sophisticated routing protocols versus mono-routing.

1.2 Plan of the thesis

We start this study by describing the concept of 3G networks, particularly the mechanisms that influence dimensioning in Chapter 2. Then, in Chapter 3, we concentrate on the studies that have already been conducted on dimensioning core and radio networks. In Chapter 4, we present the assumptions under which we worked for the traffic modeling, core and radio networks management modeling as well as the parameters to be considered for the dimensioning strategies.

A mathematical optimization model is formulated in Chapter 5. It is both an improved version and a generalization for multi-routing of a first model developed by C. Voisin in her M.Sc. thesis [1], see also [2]. In this chapter, we explain all the variables and constraints as well as the objective function. The first part of Chapter 6 describes the details of an implementation of the proposed mathematical model and the parameters used in generating the traffic instances. We then briefly introduce the CPLEX-MIP software which is used to solve the mixed linear problem corresponding to the mathematical model that has been built in Chapter 5 and discuss the options that are available to solve the mixed linear problem using a branch-and-bound method. We also present some valid inequalities that can be used to reinforce the strength of the linear relaxations of the mixed linear problem. At the end of this chapter, we validate the model using different traffic instances. While respecting the overall Grade of Service, we compare the results of multi-routing and mono-routing on different traffic instances under different assumptions on the number of divisions in the planning period and on the length of the sessions. A conclusion based on the results that have been obtained and perspectives on future work completes the thesis.

1.3 Contributions

An optimization model for the dimensioning of the 3G networks has been proposed by C. Voisin in [1], where for each requested session there is always only one path between its source and destination. In the model of [1], all potential base stations which can serve a session are assumed to be connected to the same base station controller. In addition, the soft handoff can happen only between two potential base stations. However, in practice there are always several possible routing paths between two different nodes and depending on the geographical position of each session it can be served by base stations which are connected to a unique or different base station controllers. On the other hand, when a session is accepted in soft handoff, the serving base stations are not necessarily limited to two. These arguments motivate our new developed mathematical model, generalizing [1] with the removal of the above assumptions.

In the model proposed in this thesis the routing path for each requested session will be chosen among a set of possible paths by an optimization model in order to minimize the objective function. Moreover generalizing the model proposed in [1], a session can be served by base stations connected to more than one base station controller.

Hence, in the scope of this study:

- A mathematical model which supports multi-routing in the third generation telecommunication systems is proposed.

- Different effective mechanisms for dimensioning such as, multi-routing, call admission control, Quality of Service and soft handoff between two or more base stations are considered.

- The proposed model, which provides a dimensioning tool that concentrates both on core and radio networks at the same time, is developed.

- The profit of multi-routing over mono-routing while supporting multi-services and optimal dimensioning with various traffic instances is tested, validated and evaluated.

Chapter 2

Architecture of the CDMA2000 Networks and Multimedia Services

In this chapter, we start with a general overview of CDMA (Code Division Multiple Access), the background technology of 3G networks. We go on with the description of the general concepts that will be further used and discussed in Chapter 4 and integrated in the mathematical model in Chapter 5. Indeed, in Chapter 4, we will clearly state the assumptions and the choices that will be made for the mathematical model with respect to those general concepts. Therefore, we will first discuss the concept of third generation mobile communication often called 3G as well as its requirements and services.

After talking about the 3G networks and CDMA technology, we will next focus on the CDMA2000 network architecture, which is based on CDMA technology as it corresponds to the main technology choice in North America, see [3]. In particular, we will discuss the soft handoff concept, different types of services, as well as integrated services and their specifications. At the end, some classical packet scheduling policies are presented.

2.1 Why 3G and a short history

With the appearance of data communication on the fixed network with World Wide Web (WWW), everybody expects the same ability from mobile network. That means using data services on mobile devices so we will be able to support both voice and data traffic. For a while, the Internet and mobile communications have grown separately, but they join together in 3G networks. Hence, a challenge for 3G is to bring the best features of mobile communications and the Internet together. Third generation telecommunications combine mobile radio with Internet technology to provide consumers with a new world of rich multi-media services via their mobile phones. 3G enables mobiles to carry videos, graphics and data, as well as it goes on carrying voice. It promises to deliver anytime, anywhere and anyway access for mobile users. For this reason, the existing 2G systems are replaced by the 3G systems. The first 2G systems were launched in the early 1990s with Global System for Mobile communications (GSM) and Code Division Multiple Access one (cdmaone). The evolution of the communication networks is illustrated in Figure 2.1. GSM, which is used mostly in Europe, provides a circuit-switched data service and is the most widely adopted mobile standard in the world. With over 578 million subscribers in 400 networks in 171 countries, more than 1 in 10 people on the planet used GSM technology in 2000 [4]. At the beginning the available data rate was around 9.6 kbps and then it reached 14.4 kbps. However, these range of data rates cannot support the high-speed access required for web browsing or email services. In addition, circuit-switched connections, in which a channel is dedicated to a single user, are not very efficient. GSM uses a combination of Frequency Division Multiple Access (FDMA) and Time Division Multiple Access (TDMA) to support multiple access. On the other hand, cdmaone, used in North America, is the first generation of CDMA and as its name indicates, it uses the CDMA technology to support multiple access by users. The data rate is about 14.4 kbps in this system. Next, in 2.5G networks, its data rate reached 64 kbps.

The requirements for third generation systems can be listed as below, see [5] for more details:

- Bit rates up to 2 mbps for stationary users, 384 kbps for pedestrian users and 144 kbps for vehicular users.
- Multiple simultaneous services.
- Multiplexing of services with different quality requirements on single connection, such as speech, video and packet data.
- Quality requirements from 10 % frame error rate to 10^{-6} bit error rate (core network).
- Symmetrical and asymmetrical data transmission support (radio network), e.g., web browsing causes more loading to downlink than to uplink.
- Global roaming across networks.

The 3G network has a layered architecture and is divided in two different parts (core and radio parts), which enables an efficient delivery of voice and data services. A layered network architecture, coupled with standardized open interfaces, makes it possible for the network operators to introduce and roll out new services quickly. These networks have a connectivity layer at the bottom providing support for high quality voice and data delivery. Using IP, this layer handles all data and voice connections. The layer consists of the core network equipments like routers, switches and transmission equipments. The application layer on top provides open application service interfaces enabling flexible service creation. The user application layer contains services such as e-commerce and GPS (Global Positioning System), for which the end user is willing to pay.

2.2 What is CDMA Technology?

A central base station strategy is used in all cellular networks. A link between a handset and a base station is referred to as uplink or reverse link (UL), while a link between

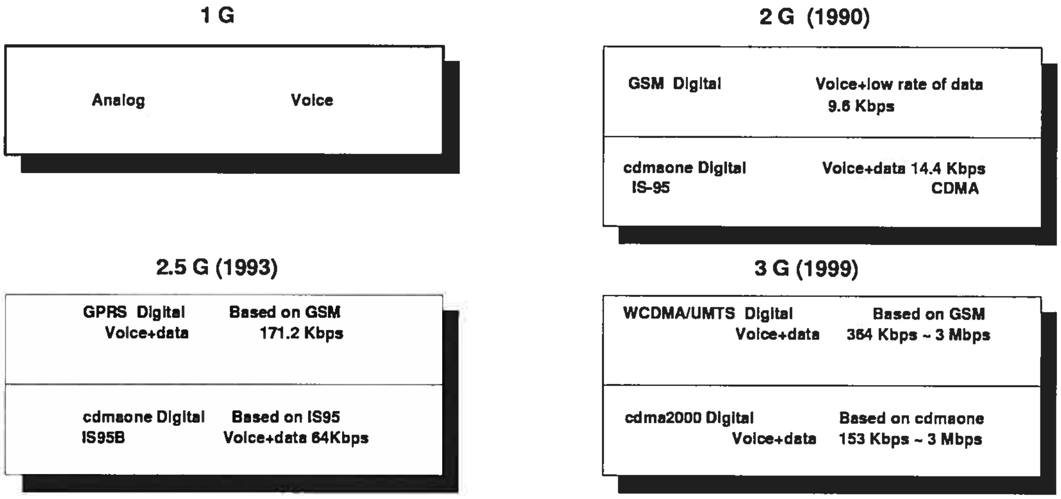

1 G
Analog
Voice
2 G (1990)
GSM Digital
Voice+low rate of data 9.6 Kbps
cdmaone Digital IS-95
Voice+data 14.4 Kbps CDMA
2.5 G (1993)
GPRS Digital Voice+data
Based on GSM 171.2 Kbps
cdmaone Digital IS95B
Based on IS95 Voice+data 64Kbps
3 G (1999)
WCDMA/UMTS Digital Voice+data
Based on GSM 384 Kbps ~ 3 Mbps
cdma2000 Digital Voice+data
Based on cdmaone 153 Kbps ~ 3 Mbps

FIG. 2.1: Evolution of the communication networks

the base station and handset is called downlink or forward link (DL). These are broadcasting channels in which each communication is assigned a unique frequency, a unique time slot or a unique code. The first is known as Frequency Division Multiple Access (FDMA), the second as Time Division Multiple Access (TDMA) while the last corresponds to the Code Division Multiple Access (CDMA), see Figure 2.2.

The cellular systems based on FDMA, such as Advanced Mobile Phone System (AMPS), have several disadvantages like the need for guard bands between signals, which reduce the available bandwidth. Meanwhile, a strong signal may capture the whole band for a long time.

TDMA offers the ability to carry data rates of 64 kbps to 120 mbps (expandable in multiples of 64 kbps), however, it is not without difficulty. Users moving from one cell to another are not assigned a time slot. Thus, if all time slots in the next cell are already occupied, a call might be disconnected. Likewise, if all the time slots in the cell in which a user happens to be in are already occupied, he will not receive a dial tone. In addition, TDMA is less robust to multi-path effects [6]. A signal coming from a tower to a handset might come from any one of several possible paths. It might have bounced off several different obstacles before arriving, which can cause interferences.

To reach a network, which is able to support wireless data services and applications such as wireless email, web browsing and digital picture taking/sending, the wireless networks are asked to do much more than a few years ago and will be asked to do more in the near future. Here, CDMA fits and provides sufficient capacity for voice and data communications allowing lots of users to connect at any given time. CDMA is the common platform on which 3G technologies are built. This technology was first used in military applications since it was difficult to jam, hard to interfere with and not easy to identify as it looks like noise.

CDMA is a spread spectrum technology and divides the radio spectrum into channels that are 1.25-MHz wide-band. Unlike FDMA and TDMA, where user signals never overlap in either the time or the frequency, CDMA allows many users to occupy the same time and frequency allocation in a given band/space (for more information see [7]). As its name implies, it assigns unique random codes to each communication to differentiate it from others in the same spectrum. The number of unique codes in CDMA is equal to the number of users. At the receiver, this code is detected and used to extract the user's information. The process of modulation of the signal by unique code is called a spreading code, spreading sequence or chip sequence.

CDMA supports two basic modes of operation: Frequency Division Duplex (FDD) and Time Division Duplex (TDD). In the FDD mode, separate carrier frequencies are used for the uplink and downlink respectively, whereas in TDD only one is time-shared between uplink and downlink. CDMA does not accept a large propagation delay between a mobile station and a base station as it causes sender-receiver collision. CDMA is considered to have numerous advantages over TDMA and FDMA. CDMA may deliver more information than FDMA and TDMA in a given time period (up to 4 to 6 times).

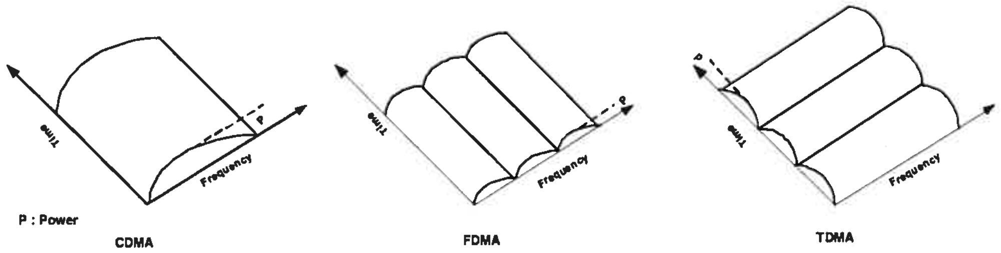


FIG. 2.2: CDMA - FDMA - TDMA

It supports soft handoff and the problem of using the same frequencies for communications within different cells (frequency reuse) does not occur in CDMA. There is also no hard limit on the number of users that we can allow on the system. Each time a user is added, the noise for the other users will be increased a little. Another advantage is that CDMA fights multi-path fading due to the fact that the signal is spread over a large bandwidth, and that each path can be tracked separately at the receiver's end [7]. CDMA is used in 2G, 2.5G and 3G networks. 2G CDMA is also called cdmaone and includes IS-95. 2.5G CDMA which is based on IS-95 is named IS-95B, while in 3G, CDMA2000 is the most famous 3G service based on CDMA.

2.2.1 CDMA2000 and UMTS

UMTS, originally developed by ETSI, is designed as an evolution from GSM toward WCDMA. The standard for this technology is developed by the 3rd Generation Partnership Project (3GPP). CWTS (China), ETSI (Europe) and TTA (Korea) are cooperating with 3GPP. The offered data rates are 144 kbps vehicular, 384 kbps pedestrian and 2 mbps when the user is not moving, see [8]. It uses the already existing GSM infrastructure. The core network of UMTS can use the current 2G networks for serving voice and packet data.

CDMA2000, developed by Qualcomm and the TIA in North America as a 3G evolution from the existing 2G CDMA system called cdmaone, originally from the IS-95 systems. The standard for this technology is developed by 3GPP2. CDMA2000 allows the simultaneous transmission of voice and data with a data rate of 2 mbps and uses the same equipments as cdmaone. It protects operator investments in existing cdmaone networks and causes a simple and cost-effective migration to 3G services. Comparing the two technologies the system capacities are more or less the same, with a little advantage of UMTS, but migration from 2G toward 3G is smoother and cost-effective through CDMA2000. Table 2.1 illustrates the technical differences between these two technologies based on [8], [5] and [9].

Air Interface Parameters	CDMA2000	UMTS
Carrier Spacing (bandwidth)	$N\times$ 1.25 MHz (N=1,3)	5 MHz
Synchronization between cell sites	Synchronous	Asynchronous
Chip rate	$N \times$ 1.2288 Mcps (N=1,3)	3.84 Mcps
Frame size	20,40 and 80 msec for physical layer	10 msec for physical layer
Modes	FDD	FDD and TDD
Multiplexing techniques	MC-CDMA	DS-WCDMA

TAB. 2.1: CDMA2000 and UMTS comparison

In both UMTS and CDMA2000 the efficient use of available resources such as bandwidth and serving different types of services are the most important issues.

2.3 Architecture of the CDMA2000 access networks

CDMA2000, which is said to have a 2 mbps data rate, is a true 3G technology based on CDMA technology. It provides higher flexibility compared to the second generation of networks. In CDMA2000 a Mobile Station (MS) can have access to a service provider network such as Internet. It includes two separate parts called Radio Access Network (RAN) and Core Network (CN).

RAN manages the radio links and soft handoff. Actually RAN is not all radio, there exists a wired part where the base stations are connected to their corresponding Base Station Controller (BSC). To understand better the functionalities and responsibilities of each part we first discuss about their elements. As shown in Figure 2.3, RAN contains Base Station (BS) and Base Station Controller (BSC), while the CN contains Packet Data Serving Node (PDSN), Mobile Switching Center (MSC), Authentication, Authorization and Accounting (AAA) and Home Agent (HA).

- Base Station (BS): Base stations are physical units of radio transmission/reception in cells. A cell is a place which is under cover of a BS. A BS has usually three antennas each covering an angle of 120 degrees. BSs can support both TDD

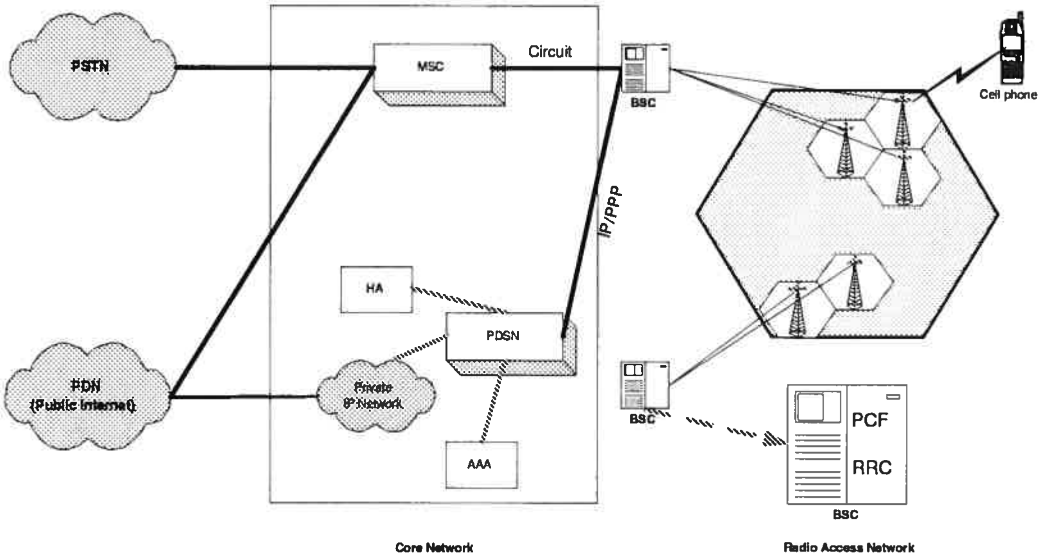


FIG. 2.3: CDMA2000

and FDD modes. They relay the calls to and from the mobile stations located in their coverage areas (cells). In other words, they provide the radio resources and maintain radio links to mobile stations. There is also a fast power control algorithm implemented in each base station. It plays a crucial role in softer handoff when the mobile station is placed in the overlap of two sectors of the same base station and asks for a session (Section 2.5.2). At this point the involved base station combines the two uplink signals received from both sectors. It is important to mention that each base station is connected to only one BSC and has all the necessary functions for its own management.

- Base Station Controller (BSC): BSCs are equipments used as interface with the core network. BSCs control the BSs as well as received and sent radio packets. They perform other radio access and link maintenance functions such as soft handoff and user mobility in a 3G wireless network. They also perform voice compression. BSCs contain two different components each with specific functionalities.

  - Packet Control Function (PCF): PCF component selects and establishes the connection to the PDSN and forwards the information to it and vice versa. In the soft handoff cases, the serving PCF sends its information to the target PCF to regenerate the packet data session to the PDSN.
  - Radio Resource Control (RRC): The RRC supports authentication and authorization of the mobile stations for their radio access.

  As well as a base station, a BSC, has all the necessary functions for its own management.

- Packet Data Serving Node (PDSN): PDSN is another component in CDMA2000 architecture and acts as a foreign agent. It performs two basic functions. It has the ability to relay the packets to the mobile stations through the Radio Access Network. Vice versa routing and relaying of the packets to the other IP networks is another responsibility of the PDSN. It provides foreign agent supports and also initiates acts as an Authentication, Authorization and Accounting (AAA) for the users. It also manages the Point to Point Protocol (PPP) with the mobile terminal.

- Mobile Switching Center (MSC): This is an interface between Public Switched Telephone Network (PSTN) and wireless system. This server is responsible for verifying the authentication and authorization of the mobile station in the RAN since it stores the authentication and authorization information for the Radio Access Network.

- Authentication, Authorization and Accounting (AAA): The AAA servers interact with the PDSNs (foreign agents) and other AAA servers to perform the functions in a secure mode. AAA provides user profile, Quality of Service (QoS) and keeps track of who, what, when and where the sessions are coming from and destined to (Accounting). The AAA server, also contains the data of users who are registered on the network.

- Home Agent (HA): It maintains user registration information and directs IP packets to the PDSN.

2.4 Core Network

The core network in CDMA2000 is based on the Mobile Internet Protocol (MIP) and connects to the Public Switched Telephone Network (PSTN) or other networks and also manages the routing. In a Mobile IP, each mobile station has a constant IP address and keeps its address even if it moves around from one point to another. This constant IP address is called home IP address. When a mobile station leaves its home network, a router named Home Agent (HA) (see Section 2.3), sends all corresponding IP datagrams to that mobile station [8].

The mobile station will use a Foreign Agent, which is a PDSN, while visiting a foreign network. It registers itself in the Foreign Agent and asks for a new address (temporary address). Then, the Foreign Agent sends that address to the Home Agent. Therefore in the case of sending a datagram to that mobile station, the HA encapsulates that in an IP packet. The destination address of this IP packet is the temporary address of the mobile station. This packet will be sent to the Foreign Agent. Then, the Foreign Agent decapsulates the IP packet and sends it to the mobile station. This method is the one which is used in the current core networks.

In both uplink and downlink choosing the proper path among the existing paths is another concern of the core network. In most networks, between all possible paths the shortest path will be chosen.

The network architecture used in this thesis contains the core network and is explained in Section 4.1.1.

2.5 Radio Access Network

In a Radio Access Network, in order to serve a request of a mobile station two steps should be considered: first, allocation of radio resources and second, establishment of a PDSN link and PPP session [10]. While a session is asked by a mobile station, a message with the required packet-data service is sent to BSC. Then the BSC sends a message to MSC in order to ask and authorize a radio traffic channel. If the answer is positive (means that the mobile station is an authorized user of the network) the BSC allocates sufficient resources for the session. Now the mobile station is authenticated and has enough resources to go on. At this point, the BSC contacts PCF (Packet Control Function) which is responsible to establish a data session with the PDSN. The PCF sends back a message based on fail or acceptance of the session to BSC. Meanwhile, a Point to Point Protocol (PPP) places between the mobile station and the PDSN to set up a packet data call. Once a mobile station has placed a PPP connection to the PDSN, it remains connected to the network.

As it can be seen in Section 4.1.1 the network architecture used in this model contains the Radio Access Network and its components.

2.5.1 Cell splitting

In order to increase the capacity of the network and decrease the co-channel interferences, the idea of cell splitting was conceived: instead of broadcasting a signal over a vast area we allow to reuse frequencies in each cell. This idea involves the base station segmentation into sectors, where there is a separate antenna for each. The most common and used cell splitting is the three sectored cell, Section 2.3. The cell is split into three sectors, with each antenna radiating a 120 degree coverage area, instead of an mono directional antenna. Each sector plays the role of a base station. There is the advantage of a stronger and clearer signal received by mobile station. Note that cell splitting is not considered in the model developed in this thesis.

2.5.2 Handoff

When a mobile station is involved in a session, it is connected to a base station via a radio link. One of the advantages of a mobile station is its mobility. To solve the problem of mobile stations getting far from the transmitting base station, handoff is introduced. Handoff occurs when the mobile station is placed in the area covered by two or more different base stations (hard/soft handoff), or when the mobile station is located in an area covered by two antenna sectors of the same base station (softer handoff).
The handoff procedure should be completed while the mobile station is in the overlap area of two or more base stations or the two antennas of a same base station. The strength of the signals and the quality decrease as the MS (Mobile Station) reaches the edge of the coverage area. The connection should be delivered to the new BS or new antenna before the disconnection of old BS or antenna from the mobile user. Otherwise the call is lost [11].

Handoff detection: Making a decision for a handoff should be based on measurements of the links at the MS and at the base station position. Since the execution of handoff costs enormously, the unnecessary handoffs should be prevented. Therefore the handoff criteria must be chosen properly. However, if the criteria are too strict, then the call may be lost before the handoff occurs [5].

Hard/Soft handoff

In this case the mobile station is located in the overlap of two or more base stations. It may happen for up to 20 or 40 percent of mobile stations. There are two possibilities in hard/soft handoff [9]:

- Micro diversity: The base stations are connected to the same BSC.

- Macro diversity: The base stations are connected to different BSCs.

In soft handoff, Figure 2.4, the communications between the base stations and the mobile station is based on one radio channel for each base station in the downlink direction. Consequently, the mobile station receives two or more signals. In the uplink direction the mobile station sends its signal. The signal will be received by base stations involved in soft handoff. Each base station sends the signal to its corresponding BSC. In micro diversity case the BSC chooses the best received signal while in macro diversity the BSCs communicate together and then choose the best. While the mobile is moving, if the mobile station leaves the overlap zone both the new and old base stations take care of the session for a certain period of time. This improves the transmission quality of wireless channel and prevents disconnection.

On the contrary, in hard handoff the link to the prior base station is terminated before or as the user is transferred to the new cell. This means, at any given time a mobile station always communicates with one base station and the old and new radio channels cannot co-exist.

Obviously, soft handoff is advantageous over hard handoff because the mobile does not loose contact with the system as it avoids interruptions and frequent switching. Nevertheless, soft handoff decreases channel availability since a mobile station may use multiple radio channels at the same time.

Softer handoff

In this case, the mobile station is located in the overlap of two sectors of the same base station, see Figure 2.5. It may happen for 5 to 10 percent of the mobile stations in a cell. Therefore, each sector uses a radio channel simultaneously. As the mobile station should be able to distinguish between the signals coming from the two sectors of a base station in the downlink direction, each signal has its unique code. Mobile station receives both signals and extracts the information. Only the soft handoff will be modeled in Section 5.4.8. The details of the soft handoff technique will be discussed in Section 4.1.3.

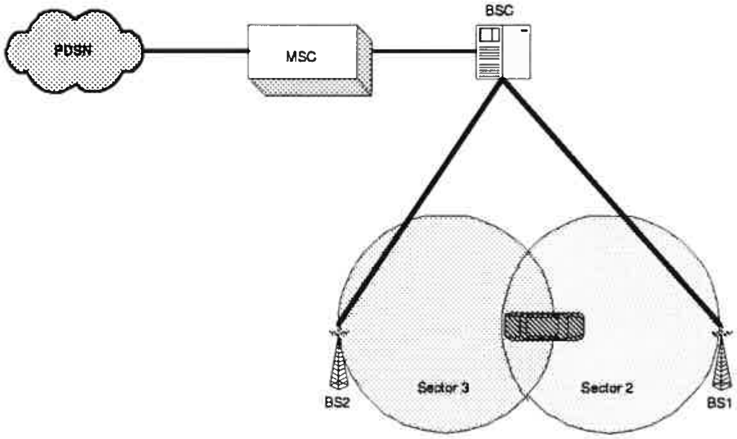


FIG. 2.4: Soft handoff

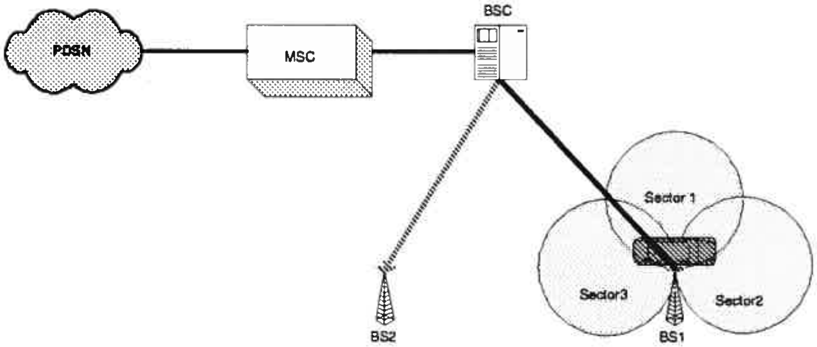


FIG. 2.5: Softer handoff

2.6 Quality of Service

With the arrival of the wireless packet-switched services in the 3G cellular networks, the dream of supporting multimedia applications, audio, video and data started to become true. One of the most important obstacles is the need for high bandwidth links. Recently, thank to improvements in coding, the need for high bandwidth is reduced and the speed of links is also increased. However, even with these changes packet-switched networks cannot fulfill the needs of multimedia as another obstacle rises which is the time of delivery.

Our concern is mostly about real-time applications such as voice and video, which are more sensitive to time as they ask for on time data arrival. Therefore, the best effort model that has been designed for non-real-time applications, and in which the network tries to deliver data but makes no promises, is not satisfying for them.

The need for a new service model where some kind of applications ask the network for higher assurance than others has become very significant. For instance, this model implies that some packets receive a particular share of bandwidth of the links or they never have delay more than a certain amount of time. QoS guarantees service requirements such as bandwidth, packet loss rate and delay. A network with these specifications is said to support Quality of Service (QoS).

We should not confuse the QoS with call service quality. The notion of service quality refers to the delivery of the service in the interaction between the user and the provider, while the QoS concerns the study of the communication services. Obviously, during the connection of two end-users we pass through several networks. At this point, the main issue is to allocate enough resources along the entire path. That is what we call End-to-End QoS.

The delay and required QoS for each type of applications are discussed more in details in Chapter 4 and Chapter 5. In this study both non-real-time and real-time applications are taken in account. Therefore, it is important to have a look at these two types of applications.

2.6.1 Non-real-time application

Non-real-time applications let the user to have an interactive communication with a server or one direction communication to another user or machine. Non-real-time applications, named also traditional data applications, like telnet and email are not so sensible to time. They can also be accepted and still be usable even with long delays and throughput is the performance goal for them.

2.6.2 Real-time application

Real-time applications such as voice and video allow communications between users on a real-time basis. Time is highly critical for those applications. It does not necessary mean that it has to be amazingly fast, it means that the tasks must be finished in a predefined time. In the real-time applications, data become digital by an analog to digital converter. Normally data are gathered at a specific rate, then they are placed in a packet and will be sent to the destination. It is at this point that the data should be played back with the appropriate rate. It seems that each part of data has a play back time. Data is useless if it arrives after its appropriate play back time. That may happen because of the delay in the network or because of possible errors and corresponding retransmission of data.

There exist different ways of dealing with this problem. One is to buffer some amount of data on receiver level. Therefore we will always have packets waiting in the buffer to be played back. It means that we have added a constant offset to the play back time of each packet, so if the packet arrives with a short delay it waits in the queue to be played back but if it arrives with a long delay it does not wait too much in the queue. Meanwhile, the offset time is much more critical for audio applications. It should not pass the 300 ms (as the partners cannot wait more than that to follow the conversation) unless the packet should be discarded. Even the real-time applications have different sensibility to loss of the data. For example, comparing audio to FTP application, loss of one bit may make the file completely wrong and useless.

The developed transport protocol by IETF to meet these requirements of real-time applications is Real-time Transport Protocol (RTP), which is rather different than Transmission Control Protocol (TCP) and with more functionality than User Datagram Protocol (UDP). It has been designed so flexible to support variety of applications, and new applications can be developed without revising this protocol.

2.7 Service Classes

The services differ in their level of QoS strictness, which describes how tightly the service can be bound by specific bandwidth, delay, jitter and loss characteristics [12].

Conversational: Conversational class is a subset of real-time applications and is strongly delay sensitive. It is always between two or more persons under the form of voice or video.

Streaming: This class of service is again a subset of real-time applications but is less delay sensitive than conversational class. It is always between a person and a data server. Transfer of data is from server to the user and can be either audio streaming or video streaming.

Interactive: Interactive class is a non real-time service where the resources are reserved dynamically. Like the streaming class it is between a person and a data server but the connection is in two directions one from human to server (request) and the other from server to human (answer). This class is delay sensitive as well as error sensitive. The transfered data should not be changed under any condition.

Background: The background class, such as mail, is a non real-time service. It is not at all delay sensitive, since the sender of the request never asks for a rapid answer in a fixed period of time. On the contrary, this service class is strongly error sensitive and data integrity is an important issue.

Table 2.2 illustrates a summary of service classes specifications. All the mentioned service classes are considered in the traffic instances used in the Chapter 6 in order to test and validate the proposed model.

Category	Characteristics	Application
Conversational	delay sensitive, real-time	Voice,Video-conference
Streaming	delay sensitive, real-time	Video
Interactive	error free, delay sensitive, non-real-time	Web-browsing
background	error free, not delay sensitive, non-real-time	E-mail

TAB. 2.2: Service classes

2.8 Integrated Services

In traditional networks, point-to-point best effort delivery was done on the model of IP. However, with the appearance of multimedia communications and real-time applications (3G in short), best effort is not an answer due to the sensitivity of the application to delay. In the context of a network with integrated services, different kinds of isolations are needed. At this point, the need for an enhanced QoS (with regard to bandwidth, packet queuing delay, routing and loss) where each individual packet asks for adequate QoS shows itself. This is what is called Integrated Services (IntServ). This Quality of Service architecture developed in the IETF around 1995-97 and often associated with Resource Reservation Protocol (RSVP). The following mechanisms are needed in order to satisfy the QoS:

- Flow specification.
- Routing.
- Resource reservation.
- Call Admission Control.
- Packet scheduling.

2.8.1 Flow specification

While sending a packet over a best effort service, we can just mention its destination. However, in the IntServ the network and different data flows need to communicate more information in terms of traffic characteristics of the flow and specifying the quality of service delivered to the flow. Thus, maybe the most important component of this architecture is the flow specification called as flow spec. This name comes from the idea that a set of packets of an application are referred as a flow and describes both the characteristics of the traffic streaming and the service requirements from the network. Describing the flows traffic characteristics is in order to give the network enough information about the bandwidth used by the flow and to let the call admission control take the right decision. The bandwidth varies for different applications and even during an application, such as video, the bandwidth is not a fixed amount.

In Chapter 6 the amount of required bandwidth in radio and wired links for all kind of applications are mentioned in Tables 6.8, 6.9, 6.10 and 6.11.

2.8.2 Routing

The decision of choosing a path from a source to a destination or destinations in case of multi-cast is called Routing. Routing is one of the most important aspects in the QoS for IntServ. The most used protocols in the current networks are Open Shortest Path First (OSPF) and Routing Information Protocol (RIP) which are based on a shortest path strategy. The shortest path can be calculated by an arbitrary metric such as number of links in a path. In a multi-service network, the priority of the sessions and their requirements are different, therefore choosing the shortest path strategy selects the shortest path for all sessions with different priorities.

According to integration of the different metrics of QoS such as delay and delay jitter the existence of an optimal routing protocol seems essential but on the other hand this protocol needs to adapt itself with the instabilities in the network. Until now there is no existing protocol which offers a compromise between stability and the coming traffic. In addition the problem is more complex when it reaches to "inter-domain" routing, where we have different administration rules and different routing policies. In this case the choice of path for a session cannot be just under the influence of its corresponding domain.

Another problem which affects the QoS and routing is the asymmetric routing. For instance the path in uplink may be different from the path in downlink and the metrics may also vary for each direction. In one it may be delay while in the other it may be delay jitter. This affects especially the real-time applications such as video and voice.
Briefly, in QoS Routing, routing associated with QoS is a mechanism in which the path for a session will be chosen both by considering the available resources and required QoS of a session. We will discuss more this topic in the next chapter.
In this study, for each requested session the serving path will be selected among a set of possible paths between the source and destination of that session. In addition, the selected path in downlink is not necessary the same which is selected in uplink. This has been discussed more in Section 4.3.

2.8.3 Resource reservation

The existing Internet Protocol (IP) in the current networks is not reliable and provides connectionless network layer services which causes the loss or duplication of packets and delays in router buffers. Since this strategy is just suitable for non-real-time applications, in the IntServ the reservation of network resources along the path helps to satisfy the required QoS for real-time applications.
An example for this kind of protocol is RSVP. White [13] has studied the RSVP and IntServ. It has been shown that RSVP can be used by end applications to select the appropriate class and QoS level. To make a resource reservation at a node, the RSVP communicates with call admission control and policy control. Admission control determines whether the node has sufficient available resources to supply the requested QoS. For this reason the admission control must consider information provided by end applications. Policy control determines whether the user has administrative permission to make the reservation. If either check fails, the RSVP program returns an error notification to the application process that originated the request.

2.8.4 Call Admission Control

Telecommunication networks aim to support IntServ over the low cost wireless services. For this reason resource sharing is considered as a major issue. Considering requests for flows with a particular level of service, the Call Admission Control (CAC) mechanism, looks at the flow specification and decides if the requested service can be satisfied with the available resources while the received QoS for previously admitted flows stay acceptable.

In other words, CAC algorithm ensures that the QoS of each connection can be maintained when a new connection is admitted. It decides whether a call can be admitted into the network based on the current traffic situation. We consider two types of CAC for the real-time and non-real-time applications.

➤ Call Admission Control in wired link

- Call admission Control for Constant Bit Rate (CBR): Since CBR is used for connections that require a constant amount of bandwidth continuously during the connection, its call admission control is not that much complicated. A CBR call will be accepted over a link if the demand bandwidth plus the current used capacity of the link does not pass the total capacity of that link. A call which is not accepted in the CAC process will be either routed again via another path or rejected. There is the possibility in which we can delay the call till the capacity of the link becomes available.

  For CBR applications the proposed CAC policy in Section 5.4.1 is based on this concept.

- Call admission Control for Variable Bit Rate (VBR): Call Admission Control for the VBR sessions is more complicated compare to CBR, since their packet rates may be different from their average rates. One way to deal with VBR sessions is to look at it as a CBR with its peak rate. Consequently, enough resources will be reserved to satisfy the session. This is the most simple solution which reduces the efficiency. There are other proposed methods which are described below:

1. Worst-case admission control: By using the results of scheduling policy we guarantee the sufficient bandwidth limit, the worst case delay and reduce the lost rate. Therefore the resources are reserved based on worst case scenarios. Since the worst case may seldom happens, it causes the inefficiency in use of resources.
2. Statistical admission control: Statistical admission control scheme obtains in advance the overflow probability when the new user will be serviced. The admission is granted if this probability is lower than the threshold which is previously set.
3. Measurement-based Admission Control: The measurement-based scheme seeks the maximum residue network bandwidth and the average residue network bandwidth through repeated measurements. These two kinds of residue bandwidths are selectively applied to the admission control through the measurements of the packet loss rate at service time. This method is very useful when we have no information about the traffic source.

As it has been explained in Section 4.1.5 the proposed CAC policy for VBR applications in this study considers an upper and lower bound for the bandwidth of each type of applications.

➤ Call Admission Control in radio link

A control system is so essential in order to balance the load on the radio network and guaranteeing the QoS of the existing sessions before accepting a new session. The admission control process will be done in BSC, since we have access to the information concerning the load of cells in this level. It determines if a base station can serve a session by calculating the radio capacity. The admission control process should be done for both downlink and uplink directions separately: a session will be accepted if both the uplink and downlink call admission controls are satisfied.

In the case of soft handoff when a mobile moves from one zone to another and being served by a new base station, the Call Admission Control process helps to reduce the loss of session and guarantees the same QoS requirements.

Based on this concept we propose a CAC policy on the radio links in Section 4.1.5 and Section 5.4.1.

2.8.5 Packet scheduling

In an IntServ network where we need to support real-time communication services, for the sake of QoS and a delivery delay bound for each packet, the packet scheduling plays an essential role. The Call Admission Control strategy also depends on the scheduling policy. Being more general, a scheduling policy influences the performance that a guaranteed-service receives along the path from the source to the destination. In different policies different bandwidth will be allocated to each application. It also affects the loss rates by considering more or less buffers for different coming sessions.

Depending on the network situation and applications requirements a scheduling policy may concentrate on one of the following items [16]:

- ➢ Easy and efficient admission control.
- ➢ Easy implementation of managing the buffers and queues. Since a packet scheduling policy concerns each packet it cannot be too complex.
- ➢ Deterministic and probabilistic guaranteed-performance per session, note that the first one needs more network resource reservation. There are four main parameters for performance: bandwidth, delay, delay jitter and loss.
- ➢ Protection and fairness for current and coming sessions.

There are other fundamental issues in the scheduling disciplines such as non-work-conserving or work-conserving and priority levels. A work-conserving discipline is never idle when packets await service while a non-work-conserving discipline may be idle even when packets await service. The non-work-conserving discipline delays the packets and will make the coming traffic more predictable. Therefore, it will reduce the delay-jitter and the necessary buffer size but the implementation cost may be the biggest problem. Each connection has a priority level. Packet is served from a given priority level only if no packets exist at higher levels. Obviously, the connection with highest priority level gets the lowest delay. Since the high level packets may always exist there is the possibility of appearance the starvation where the scheduler may never answer to a packet with low level priority.

In current networks, all packets are served on a best-effort, First-Come-First-Served (FCFS) basis. This method is implemented using FIFO (First In First Out) queue (add to tail, take from head) and is a work-conserving discipline. Incoming packets are in order in the queue and a packet will be lost when the queue is already full. The scheduler cannot distinguish the sessions therefore the QoS requirements for each session cannot be satisfied. Simplicity is its main advantage, so it is easy to implement and requires few resources.

For the best-effort connections where we need a max-min fair allocation, General Processor Sharing (GPS), a work-conserving discipline, is introduced. The packets are placed in separate logical queues. Since GPS visits queues once in an interval, each queue has its chance to send its packet on the network. If one queue has nothing to send, it is skipped and the saved time is divided between other queues. This method offers protection, but is not at all easy to implement.

Weighted Fair Queuing (WFQ) is used both for best-effort and guaranteed services and is also a work-conserving discipline. Its aim is to let several sessions share the same link. WFQ is equivalent to Packet-by-Packet GPS (PGPS). It first computes the time at which a packet will complete service using GPS and then serves packets in the order of this time. The current round number and the highest per queue finish number are two important concepts in this method. The round number is the number of rounds of service completed by a bit-by-bit round robin scheduler at a given time and may not be always an integer. By knowing the round number we can calculate the finish number. In an inactive connection the finish number of a packet is the current round number plus the packet size in bits, while in an active connection it is equal to the sum of the largest finish number of a packet in its queue and the size of the packet in bits. Comparing the last two methods a connection in WFQ can receive more services than in GPS. In addition WFQ is fair and provides real-time performance guarantees. It also performs well with variable size of packets and does not need to know the average packet size in advance. On the other hand, it is complicated to implement because of its pre-connection requirements and it also suffers from iterated deletion.

	FCFS	GPS	WFQ	WF2Q	D-EDD	J-EDD	RC
Work-Conserving	√	√	√	√	√	-	√
Non-Work-Conserving	-	-	-	-	-	√	√
Best-effort Service	√	√	√	√	-	-	-
Guaranteed Service	-	-	√	√	√	√	√
Bandwidth	-	-	√	√	√	√	√
Delay Bound	-	-	√	√	√	√	√
Delay Jitter	-	-	-	-	-	√	√

TAB. 2.3: Comparison table for the scheduling policies

WF2Q is called Worst-case Fair Weighted Fair Queuing with a work-conserving discipline. In this method only packets who have a virtual start time that has been passed are considered for output. It is again another approximation of GPS and has lower delay bounds but needs even more complicated implementation compare to WFQ.

In the Delay-Earliest-Due-Date (D-EDD) scheduling which is a work-conserving discipline, a deadline is assigned to each packet. It assigns scheduling deadlines so that even with all connections at peak rate, worst-case delay in traffic descriptor is met. Packets are served in order of their deadlines and the admission control makes sure that all deadlines can always be met. Its advantage over WFQ is that for each session it provides end-to-end delay bounds independent of the guaranteed bandwidth of that session. While in a Delay-Jitter-Earliest-Due-Date (J-EDD) scheduling, which is a non-work-conserving discipline, all packets receive the same delay at every hop except at the last hop. In order to reduce delay jitter, all packets receive a large delay. This method can provide end-to-end bandwidth, delay and jitter bounds.

The last scheduling policy discussed here is Rate Controlled (RC) scheduling in which the bandwidth, delay and delay-jitter bounds are provided. It can be a work-conserving and non-work-conserving discipline. The packet will first be placed in the regulator and after calculating their eligibility time they will be sent to scheduler. At this level the scheduler selects among eligible packets to send over the network. The last two methods are suitable for the real-time applications as they bound the delay jitter and consequently the size of buffer in the destinations. On the contrary, they are both complex to implement. Table 2.3 illustrates a comparison between the mentioned scheduling policies (for more information see [16]).

In order to calculate the queuing delay in each node we consider the WFQ scheduling policy. This has been discussed in Section 4.4.1.

Chapter 3

Literature Review

Different methods and optimization models have been proposed and studied for the network dimensioning problem. In most of these studies the authors aim at minimizing the cost of services while providing reliable connections that satisfy as much as possible the users of the network, i.e. the Quality of Service (QoS) constraints. As mentioned in the previous chapter, a 3G network is a combination of a core network and a radio network that communicate together. Hence, both core and radio networks should be considered in the dimensioning in order to minimize the number of equipments and satisfy the quality of service at the same time.

An overview of the papers studied on the core network dimensioning specially multi-service IP network and radio network dimensioning can be found in this chapter. In most of these papers the core and radio networks are studied individually, therefore we divide this chapter into core network dimensioning, radio network dimensioning and core and radio network dimensioning.

3.1 Core network dimensioning

OSPF (Open Shortest Path First), RIP (Routing Information Protocol) and BGP (Border Gateway Protocol) are called best effort routing protocols and are currently used in Internet [14]. They use only the shortest path to the destination, while shortest path here does not necessarily mean the path with the shortest physical distance. It may also mean the path with the least cost or fewest hop counts. Current protocols use single objective optimization algorithms which consider only one metric (bandwidth, hop count, cost). Thus, all the traffic is routed on the shortest path, even if there exist some alternate paths. The alternate paths are not used as long as they are not the shortest ones. The disadvantage is that it may lead to congestion on some links, while some other links are not fully used.

QoS routing is supposed to solve or avoid the problems mentioned above. It is the process of selecting a path to be used by a flow based on QoS requirements for multimedia applications with the efficient use of resources. Obviously, this is not an easy task as services have different QoS constraints. When there are several feasible paths available, the path selection can be based on some policy constraints. The motivation of path selection is to improve the service received by users and the network dimensioning.

There are many proposed QoS-based routing algorithms which are mostly based on the current best effort routing strategy. That is because these two routing strategies (best effort and QoS-based routing) must be able to coexist. The routing protocols which are currently used in Internet are Distance Vector and Link-State algorithms [14]. While working on the QoS routing we always take in account that the new algorithms should be suitable for the existing Internet architecture, efficient and scalable to a large network and easy to implement, besides not too complex in order to be efficient in a real-time environment.

We can classify the quality of service based routing algorithms into 3 strategies [17],

[18].

1. **Source Routing**

   In this strategy, each node has a complete information about the network topology and the state of every link. The best path will be calculated based on this information. This method is easy to implement and avoids dealing with distributed computing problems. Everything is decided at the source level and routers on the way follow the pre-defined path. Note that the information should be updated frequently at each node.

2. **Distributed Routing (hop-by-hop routing)**

   The path is computed in a distributed fashion. Each router only knows the next hop toward the destination node. Control messages are exchanged among the nodes and the information in each node helps to choose the best path. Thus, when a packet arrives at a given node, the router sends it to the next hop. This method is also called hop-by-hop routing. This is used by most current "best effort" routing protocols such as RIP. It decreases the routing time as the routing computation is distributed among all routers on the way to destination. However, when the routing state information in different routers is not consistent, for instance the information is updated in a router but not in another, it may cause some routing loop problems.

3. **Hierarchal Routing**

   In this strategy, groups of adjacent nodes are defined and each group is a logical node in the higher level group. Every node maintains an aggregated global state, which contains the state information of all nodes within the group and the information about the other groups. Thus, the routing computation is shared by many nodes. As a logical node (a group) may contain a large subnet with a complex structure, this may have a significant negative impact on QoS Routing. The difficulty increases when multiple QoS constraints have to be taken in account.

A performance comparison [17] between different studied algorithms is reported in Table 4.2, where v is the number of nodes and e is the number of links. The word "generic" means that a routing framework is proposed, from which we can derive different QoS constraints.

Routing Strategy	Algorithm	QoS Constraints	Routing complexity
Source	Ma-Steenkiste [19]	Bandwidth Constraint	O(1)
	Gurin-Orda [20]	Bandwidth Constraint	O(1)
		Delay Constraint	O(1)
	Wang-Crowcroft [21]	Bandwidth, Delay Constraint	O(1)
	Chen-Nahrstedt [22]	Bandwidth Cost Constraint	O(1)
Distributed	Cidon et al [23]	Generic	$O(e)$
	Salama at al [24]	Delay Constraint, Least Cost	$O(v^3)$
	Wang-Crowcroft [21]	Bandwidth Constraint	$O(v)$
	Chen-Nahrstedt [25]	Generic	$O(e)$
Hierarchical	PNNI [26]	Generic	$O(v)$

TAB. 3.1: Comparison table

In all cited algorithms, a global state needs to be maintained at every single node. In addition most of them have considered the routing problem as a shortest path problem and tried to solve it by Dijkstra or Bellman-Ford algorithm.

In [27] the authors have considered network dimensioning and performance of multiservice in a 10 node network with dynamic routing. They have assigned different peak rates to different services while each service has a fixed peak rate during the whole session. In this study, the length of a path is limited to a maximum of 2 links, therefore a session can be routed from a source to a destination either on direct path or can pass through at most one intermediate node.

The objective is to minimize:

$$\sum_{\ell \in L} c_\ell y_\ell$$

where:

L Set of traffic links,

c_ℓ Cost of capacity unit on the traffic link ℓ,

y_ℓ Required capacity unit on traffic link ℓ.

In comparison with the previous works done on routing which were mostly on single service cases, this is an evolution. Although, GoS and call admission control are considered, the study suffers from the limitation in path length, small size of the network and lack of QoS constraints.

Medhi and Sukiman in [28] worked on multi-service (2 services) dynamic QoS routing in a circuit communication. A peak rate allocation of bandwidth during a session which concerns the Constant Bit Rate (CBR) sessions associated with GoS and call admission control mechanisms are considered in this paper. The bandwidth unit is called BBU (Basic Bandwidth Unit) and the number of BBUs for each kind of connection is named SU (Service Unit). Even in this study the restriction on at most 2 links for each path exists. In this paper an equitable GoS for all services is assumed, e.g., 1% blocking for all services.

In this model each session should go through the two following phases:

- Call admission control phase,
- Routing phase.

The proposed call admission control is based on the available free capacity on the direct link if there is any and follows a probabilistic decision scheme. Considering a call for a service type s arrives for the node pair i, j: this call is then accepted (not connected yet) to the network with a certain probability. If the call is accepted in the admission control phase, then it goes to the routing phase to determine if the call can be actually routed using any of the following routing schemes. The admission control can be given for service $s \in S$ for traffic pair i, j by the following acceptance function:

$$\alpha^s_{[i,j]} = \begin{cases} P^s_{[i,j]} & \text{if } L(i,j) \leq t(i,j) - a(i,j) < U(i,j) \\ 1 & \text{otherwise.} \end{cases}$$

where $t(i,j)$ is equal to the total number of BBUs on link (i,j), $a(i,j)$ is equal to the number of BBUs on link (i,j) that are presently allocated to active calls of all types, $L(i,j)$ and $U(i,j)$ are lower and upper bound, respectively, on free capacity for probabilistic acceptance and $0 \leq P^s_{[i,j]} \leq 1$. If $P^s_{[i,j]} = 1$, $s \in S$ for all traffic pairs, then no admission control is in the network. Such a call admission control is local and is not based on a complete network information.

In [28] different routing schemes are studied:

1. **Max available capacity routing with periodic update**

   For each session between each pair of source and destination, the session will be accepted if there is enough capacity on the direct path, otherwise the first alternative path via existing node in the routing table will be examined. If none of the alternative paths can serve the session, the session is blocked. The necessary computation can be done in a centralized manner where each switch sends its updated informations to the central processor in a distributed manner using link-state management to send the residual capacity information.

2. **One-call old routing**

   The idea is again to try the direct path first, and only one alternative path is stored. If the direct path is not able to serve the session the stored path will be tried and meanwhile a new alternative path will be computed. The goal of this method is to minimize the call setup time for calls. This is done on a per call basis and the information available from the last call is first used for the newly coming call.

3. **Force one-call old routing**

   This method is similar to the previous one with less computation. Some service use one-call for alternate routing while others use freshly computed.

4. **Probabilistic one-call old routing**

   On the arriving of a session, the direct link is examined first. If the available bandwidth is not sufficient on the direct link then the decision for routing this call will be made based on the information about the routes available from last call.

Only 2 services, voice and video, are considered in a 10 node network with a fixed GoS for all kind of services and the routing paths have at most 2 links. These computations that are proposed could be quite heavy in the case of larger networks and longer routing paths.

In [29] the QoS requirements are limited to the number of hops (or cost-based path requirements) and services that require bandwidth guarantee. There are three steps to follow:

- ➢ A set of possible paths and their storage will be considered.

- ➢ This set of paths will be filtered in order to provide a set of QoS acceptable paths. The order of the routes in the set may change from most acceptable to least acceptable, e.g., based on path residual bandwidth.

- ➢ The best path will be selected among the sorted paths.

Three types of methods are examined; in one, which supports crankback (this concept allows a session which is in process to return to the source and try another path), the direct path to the destination is checked first, i.e. the residual bandwidth on the path is evaluated. If there does not exist any direct path, another path will be considered.
In the other method, there is no update and the information is what it was on the beginning. In the third method after doing the first step the whole system tries to find the path with the most available capacity which causes a high load on the network.
It has been experienced that in a weakly loaded network with heterogeneous services, inaccuracy in state informations can be compensated by crankback system. Although, the crankback does not increase flow setup time in a weakly loaded network it can be a problem in a moderate or high-load network therefore having exact information in real time leads to an increase of the flow set-up time. In addition, in this study the end-to-end delay requirements and packet loss rate are not included in the QoS routing computation.
Jiang and Papavassilious [30] propose an Optimal Least Weight Routing (OLWR) dynamic algorithm for QoS routing in high speed networks. This algorithm chooses the best and optimal path based on calculating the minimum weight parameters for each available path in the set of candidates.
A uniform distribution of traffic in the entire network has been considered in this paper. The proposed algorithm has the following steps:

- Traffic classification before entering the network.
- Estimation and calculation of effective bandwidth.
- Choosing the optimal path by considering the effective bandwidth and networks model.

The performance evaluation is done using the OPNET modeling and simulation tool and the bandwidths of all links are assumed to be equal. Each session arrives in the network according with a Poisson process of rate λ_i.

With:

b_i^h Effective bandwidth in route R for class i with h hops,

φ_i Departure rate of class i,

λ_i Arrival rate of class i,

B_i Blocking probability of class i.

the expected revenue and the carried load (actual traffic that is accepted by the network) are defined respectively as follows:

$$\text{Revenue} = \sum_{i=1}^{n} \frac{b_i^{min(h)}}{\varphi_i} \lambda_i (1 - B_i), \tag{3.1}$$

$$\text{Carried load} = \sum_{i=1}^{n} \frac{\lambda_i}{\varphi_i} (1 - B_i). \tag{3.2}$$

In this model the higher the class, the larger effective bandwidth it requires, therefore $b_1^h \leq b_2^h \leq ... \leq b_n^h$.

This algorithm tries to prevent the acceptance of low priority sessions affect the high priority sessions and allows routes with more available bandwidth for higher bandwidth traffic. Within a simulation with 10 calls per second OLWR seems to have an acceptable blocking probability of the highest traffic class but a very high blocking probability for lower service classes.

A routing protocol which is not constrained by weight and can be implemented with modifications in address prefix forwarding mechanism is studied in [31]. This protocol is based on flow optimization but referring to the article itself “this approach is perhaps not realistic when it comes to deployment in real networks.” The objective is to minimize:

$$f(y) = \sum_{t,(i,j)} b_{ij}^t y_{ij}^t, \tag{3.3}$$

where:

b_{ij}^t Cost on the link (i, j) to the destination t,

y_{ij}^t Traffic to the destination node t routed through the link (i, j).

3.2 Radio network dimensioning

There are two main directions along which we can dimension a radio network: Area Coverage and Demand Coverage. The first one focus on providing the best radio signal for each location of a specified region. This dimensioning objective has been studied in [32] where the authors consider the minimization of the number of base stations and their optimal geographical position with a non-linear programming model. The authors include constraints in order to make sure whether the Signal to Interference Ratio (SIR) received by a session from the base station is sufficient for a signal to be understood clearly. If the existing base stations are not able to satisfy the coming sessions then a new base station will be added to the system.

Demand coverage approach appeared to handle the growing size of radio networks. Therefore both capacity and needed equipments became crucial to reduce the service cost. [33] studies such approach and considers the density of users on a specific region. Assuming a fix number of base stations, the objective is to maximize the proportion of demand nodes covered by the cells within the permitted range. The authors introduce the concept of a demand node that represents the center of an area that contains a quantum of demand accounted in a fixed number of call requests per unit of time. This leads to the following formulation:

$$\max \sum_{j \in J} a_j y_j, \tag{3.4}$$

where:

I	The set of potential positions for the base stations,
J	The set of demand points,
p	Number of base stations,
a_j	Population at demand node j,
$N_j = \{i \mid f_{ji} \leq PL\}$	The set of positions i for the base stations while the attenuation of signal f_{ji} between i and the demand point j has a value smaller than the limit value PL which ensures a signal with enough power on the receiver.

$$x_i = \begin{cases} 1 & \text{if a base station is selected in position } i. \\ 0 & \text{otherwise.} \end{cases}$$

$$y_j = \begin{cases} 1 & \text{if the demand point } j \text{ is covered by a base station.} \\ 0 & \text{otherwise.} \end{cases}$$

with the following constraints:

$$\sum_{i \in N_j} x_i \geq y_j \qquad j \in J, \tag{3.5}$$

$$\sum_{i \in I} x_i = p \qquad j \in J. \tag{3.6}$$

Constraint (3.5) implies that the demand node can be covered if and only if one base station is located within the standard range. This means that if a demand point is covered, then there is at least one base station which guarantees a minimum SIR. Constraint (3.6) forces the number of placed base stations to be exactly equal to p. The coefficient a_j is used to consider the different priorities for different demand points. For instance, airport can be considered as a traffic point with higher priority.

As seen above all these studies are restricted to the selection of base stations location and all the base stations are assumed to give service to the traffic simultaneously. On the other hand, the capacity of base stations are not considered explicitly. On this last point Lee et al. in [34] worked on cell planning in radio network where two types of base stations existed. Some base stations are currently in service (active) and some are new, ready to give the service and be active. The objective of this model is to minimize the cost of new base stations

$$\min \sum_{k=K_1+1}^{K_1+K_2} c_k z_k, \tag{3.7}$$

where K_1 is the number of already existing base stations, K_2 is the number of added base stations, z_k is equal to one if the BS_k is selected and c_k is the cost of base station BS_k. A tabu search algorithm is used to solve this model which is formulated as an integer linear problem. Considering the existence of 2500 Traffic Demand Area (TDAs) in CDMA, the tabu search has been able to reduce the cost around 10 to 20 %. Note that potential service area of a base station represents the TDAs that can be served with sufficient quality by that base station.

3.3 Core network and radio network dimensioning

None of the papers studied in 3.1 and 3.2 has considered both radio and core network dimensioning. They have concentrated either on core or on radio network. While in [1] dimensioning in both core and radio networks and their impacts on each other has been studied. The objective is to minimize the used capacity of wired links in both core and radio networks and the number of activated base stations. The soft handoff aspect as well as Call Admission Control (CAC), Grade of Service (GoS) and Quality of Service (QoS) have been considered in this model. However, a static routing strategy with choosing a random single path for each pair of origin and destination limits the dimensioning of the core network. In radio network part, a session in soft handoff is assumed to be served by two base stations. All the potential base stations that can serve a session are connected to a unique BSC, which is another limitation of this model. The proposed mathematical model in the current M.Sc. is inspired from the model discussed in [1].

3.4 Conclusion

The real needs of multi-service 3G networks are not considered in the previous studies. In the core network, most of the works do not consider multi-services. Moreover they suffer from a small core network or considering just one path or maximum 2 links for each path. These assumptions are not realistic for the 3G networks. In the models where the multi-service has been considered, the delay constraint for Quality of Service is not taken into account, although delay is an important issue for the real-time applications.

Having a look at what has been done on the dimensioning of radio network, very few studies have considered multi-service. Most of the studied models have only considered the voice service. Some papers have focused on covered surface of cells and the interferences. While these are interesting for the dimensioning, they do not include guarantees for data rates and Quality of Service for each requested session. The needs of 3G networks lead us to work more on the dimensioning tools which considers different data rates for different applications. Restricted assumptions made for the analytical expressions of the capacity of the base stations while considering individual Quality of Service conditions for each user are the obstacles in this direction of study.

In most of studies done on dimensioning the multi-service 3G networks the core and radio networks are studied individually. While the most challenging part of a global dimensioning model for the core and radio networks is to consider them together with the bottleneck issue of the downlinks in the context of multi-service.

Chapter 4

Dimensioning Strategy and Network Modeling

In the previous chapters, we presented an overview of the CDMA2000 technology for the 3G networks and its specifications. We discussed some of the works and studies performed on 3G networks more specifically with respect to the Quality of Service and to Routing. The aim of the research project of this M.Sc. is to analyze the impacts of multi-routing on the dimensioning of a 3G network when we support multi-services. While talking about dimensioning, we focus on both radio access network and core network. Our objective is to minimize the number of base stations in the radio access network and the capacity of the wired links in both the core network and the radio network. In this section, we present the network model, the dimensioning aspects and the characteristics of the traffic model.

4.1 Network modeling

In the modeling of a network based on packet communication, we include a Call Admission Control (CAC) procedure and Grade of Service (GoS) conditions. CDMA2000 includes the following layers of the OSI model:

- The physical Layer which corresponds to the OSI Physical Layer.
- The medium Access Control (MAC) which corresponds to the OSI Data Link Layer.
- The link Access Control (LAC) which corresponds to the OSI Data Link Layer.
- The signaling, packet and data service which corresponds to the OSI Network Layer and up.

In both the core and radio networks there are some protocols and technologies which play key roles in dimensioning. In this model, for core network the IP and RSVP protocols for IntServ and the WFQ scheduling policy have been assumed. In radio network the Selective Repeat Protocol (SRP) and the concepts of CDMA are used.

4.1.1 3G network architecture

Let us recall the main features of the radio and core networks that are taken into account in the mathematical model. Figure 4.1 summarizes those features.

- ➤ Core Network (CN):

  In 3G networks, the CN part provides the links between the radio network and the external networks. A telephony network or the Internet are examples of external networks. A core network is defined by both a set of nodes and a set of wired links. Nodes of the CN and their responsibilities are discussed in Section 2.3. All nodes are considered "equal" in our model except for the external nodes which play the role of destinations for sessions.

- ➤ Radio Network:

  The radio network includes the mobile stations which play the role of origin nodes for sessions, the base stations which are connected to BSCs (each base station is connected to only one BSC) and a set of radio links and wired links. The

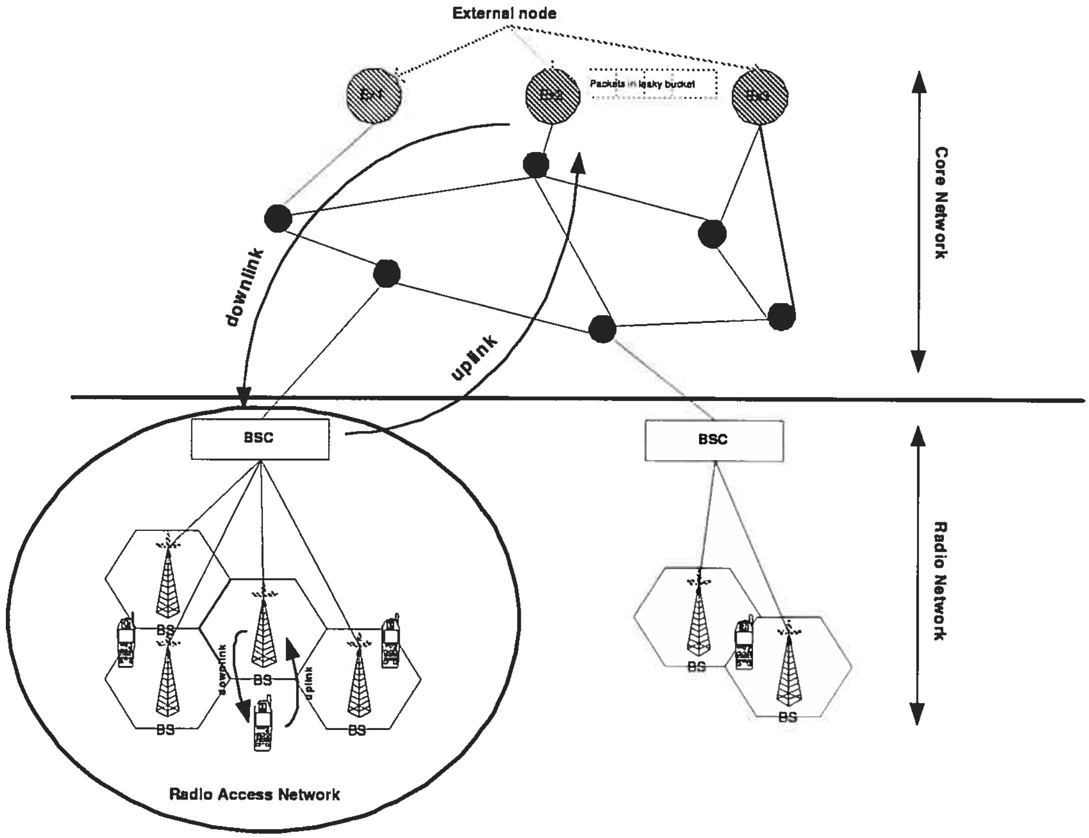


FIG. 4.1: Network model

connections between mobile stations and base stations are of radio type, while the base stations are connected to their corresponding BSC by wired links. The BSCs are located at the frontier of the radio network and core network and lead the sessions to/from base stations and mobile stations to/from the destination, which is an external node, through the internal nodes. In each BSC there is a built-in table which contains the list of paths to different external nodes. They also manage the resources and handle the soft handoff process. A Radio Access Network (RAN) is defined by a BSC and the set of base stations connected to this BSC.

An uplink direction defines the connection of a mobile station toward an external node and a downlink defines the connection of an external node to a mobile station.

4.1.2 Distribution of the BSCs and base stations

Each network instance is characterized by width and length of a geographical area, a number of BSCs, a number of potential base stations and the wired links (network topology). We also know the number of possible sessions requested for a planning period. First the geographical area is divided between the BSCs, therefore each one has its own covering area. The potential base stations are uniformly distributed on that area. Depending on the position of each base station we can find out its corresponding BSC. Assuming that each base station is connected to only one BSC, the BSC associated with a base station can be found easily. If a BS is located in the covering area of a BSC then that one is its corresponding BSC, Figure 4.2. We call them potential base stations as not all of them will be selected in the solution schema corresponding to the minimum dimensioning cost.

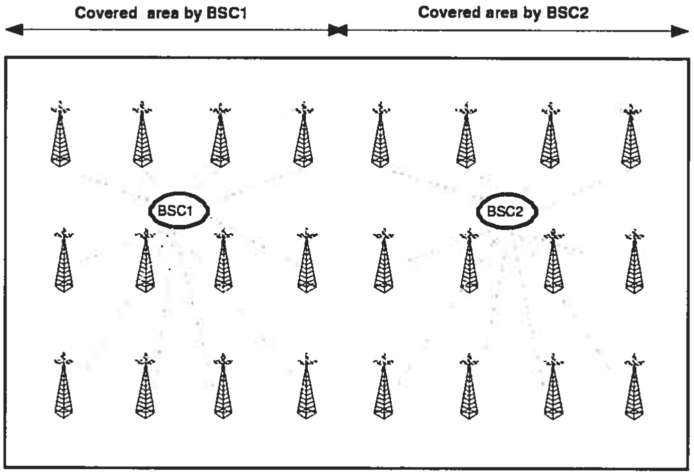


FIG. 4.2: Distribution of BSCs and BSs

4.1.3 Soft handoff

As discussed in Section 4.1.2, the base stations are assumed to be uniformly distributed and each of them is connected to a unique BSC. In this model, for each session we distinguish two sets of potential base stations. One contains the base stations which can serve the session only in soft handoff and the other is a set of base stations which can give service to that session both in soft handoff and non soft handoff. These selections are based on the distance of the session from each base station. Mobile stations are assumed to be fixed (no mobility). It means that the geographical position of the mobile station is fixed during the life time of a session. Cell splitting, Section 2.5.1, has not been considered, therefore only soft handoff is modeled, not softer handoff. As mentioned before, soft handoff can happen between two or more base stations which are connected to the same BSC. However, we do not limit a session to one BSC. Each potential base station for a session can be connected to a different BSC. Recall from Section 2.5.2 in the soft handoff process that radio links between mobile station and base stations work in parallel. Base stations involved in soft handoff provide the same service for each of the flows with the same bandwidth and frame error rate. In the uplink direction, the BSC receives the frames coming from two or more base stations in soft handoff and chooses the best among them to send over the core network. However, in the downlink direction the task of choosing the best frame among the frames coming from different base stations that are involved is done at the level of the mobile stations. The impact of soft handoff on dimensioning is not deniable. A session in soft handoff occupies the radio capacity on more than one base station. Note that when there is a soft handoff, the power of the signals which are sent is much weaker comparing to those sent when there is no soft handoff. This comes from the multiplicity of the radio links and the action of comparing the received signals in the BSC on the uplink and in the mobile station on the downlink for choosing the best signal. Consider a mobile station located at the end corner of the covered zone of a base station which asks for a connection. It has to send a very powerful signal on the radio link toward the base station in order to have an acceptable quality. However, if more than one base station serve this session, it needs a less powerful signal to be sent for the same quality.

Soft handoff has an impact on the radio link capacity on both the downlink and uplink directions which we call soft handoff profit. Indeed the required ratio of signal energy per bit to noise spectral density (E_b/N_t) for a mobile station in soft handoff is less comparing to non soft handoff. Note that a weak E_b/N_t does not let the signal to be correctly decoded in the receiver, Section 4.1.4.

4.1.4 Capacity of the radio links

We use the capacity formulas that are established in [5], [35] and [1]. In [5] the formulas are based on WCDMA which use the CDMA but with a frequency bound of 5 MHz. Moreover, In CDMA2000 the maximum frequency bound is 1.25 MHz. The formulas of [5] remain valid with some changes in numerical values. Moreover, in this study we have generalized the soft handoff and allow it not only between two, but also between three or more base stations. The rate at which the data can be transmitted over a given communication path, or channel, under given conditions, is referred to as channel capacity. There is a capacity threshold that determines on whether to accept or reject a coming session. The calculation is done in two steps: first, without taking in account the soft handoff and second with the soft handoff profit. At this point we need to clarify and explain some concepts for better understanding of the capacity formulas.

Radio Access Bearer (RAB)

Radio Access Bearer, RAB, is defined by two parameters, a discrete value for the data rate and the Frame Error Rate (FER). If we denote a RAB by r then r_t corresponds to data rate and r_{fer} corresponds to its FER. The available RABs on the radio links are not always the same. We may use different RABs for different types of applications. For instance, a voice session which does not need a high data rate may use a RAB r with low r_t but less r_{fer}. There exist different combinations of RABs. We use the combinations described in Table 4.1, called $RC4$ according to [36].

$(\frac{E_b}{N_t})$ Ratio of signal energy per bit to noise spectral density

The ratio of signal energy per bit received at the base station to noise spectral density per hertz , $\frac{E_b}{N_t}$, should be powerful enough so the signal can be decoded and therefore

Data Rate bps	FER (%)
9600	0.5 , 1 , 2 , 5 , 10
19200	0.5 , 1 , 2 , 5 , 10
38400	0.5 , 1 , 2 , 5 , 10
76800	0.5 , 1 , 2 , 5 , 10
153600	0.5 , 1 , 2 , 5 , 10
307200	0.5 , 1 , 2 , 5 , 10

TAB. 4.1: RAB combinations in *RC*4

understandable by the receiver. It is measured at the input of the receiver and is used as the basic measure of the signal strength. In the uplink direction this ratio will be checked by the base station while in the downlink this will be done by the mobile station. If this ratio of the signal has a small value, then the signal decoding is not going to be possible. Here noise means the sum of thermal noise power spectral density, N_0, and interferences. A value for ratio of signal energy per bit to noise is assigned to each RAB. This ratio is equal to $\frac{S}{N}$ (Signal-to-Noise) ratio only when the bandwidth is equal to the data rate. $\frac{S}{N}$ ratio depends on bandwidth and data rate but $\frac{E_b}{N_t}$ is independent of bandwidth and data rate. The bit error rate for digital data is a function of $\frac{E_b}{N_t}$. As the bit rate increases, transmitted signal power must increase to maintain a required $\frac{E_b}{N_t}$, an increase in data rate increases the error rate. The relation between this ratio and the power of the signal can be seen in the following formula:

$$(\frac{E_b}{N_t})_{s,r} = \text{SIR}_s \times \frac{W}{r_t} \tag{4.1}$$

where SIR_s is the ratio of signal to noise and is equal to $\frac{\text{power of received signal for the session s}}{\text{power of thermal noise and the interferences}}$, W is the spreading bandwidth and r_t is the data rate of selected RAB for the session s.

Attenuation factor

For each signal the attenuation corresponds to the loss of energy during its propagation. Therefore, there is always a difference between the power of sent and received signals.

The relation between these two signals, see [35], is as follows:

$$P_r = P_s(\frac{h_s h_r}{d^2})^2, \tag{4.2}$$

where:

- P_s : Sent Power.
- P_r : Received power.
- h_s : The height of sender.
- h_r : The height of receiver.
- d : The distance between sender and receiver.

For a session s which is being served by a base station BS_0, the attenuation factor, see [35], can be calculated as follows:

$$\text{Attenuation Factor} = \frac{\text{Sent Power}}{\text{Received Power}}. \tag{4.3}$$

From the equations (4.2) and (4.3) we have:

$$\text{Attenuation Factor} = \frac{d^4}{h_r^2 h_s^2}. \tag{4.4}$$

In this model the attenuation factor is explicitly considered in the downlink direction. We consider that the attenuation factor is already taken into account in the powers received by the base station in the uplink (P_s^{UL} and P_T^{UL}) direction.

Notations

The following notations are used in the calculation of a given base station capacity.

- S : The set of sessions taken by the base station.
- W : Bandwidth value for CDMA2000.
- $\gamma_{s,r}^{\text{UL}}$: $(\frac{E_b}{N_t})$ associated to a session s served by RAB r in uplink.

- $\gamma^{\text{DL}}_{s,r}$: $(\frac{E_b}{N_t})$ associated to a session s served by RAB r in downlink.
- λ : Inter-cell interferences which is a factor λ of intra-cell interferences.
- SHG^{UL} : The reduced factor of $(\frac{E_b}{N_t})$ when the session is in soft handoff in uplink.
- SHG^{DL} : The reduced factor of $(\frac{E_b}{N_t})$ when the session is in soft handoff in downlink.
- LF_{UL} : Limit load factor of a radio link in uplink.
- P^{UL}_s : The assigned power to the session s by the base station.
- P^{DL}_s : The sent power to the session s by the base station.
- P^{UL}_T : Total power received by the base station sent by the mobile stations attached to the base station.
- P^{DL}_T : Total power sent by the base station to the attached mobile station.
- P_{cont} : The assigned power to synchronization and control signals.
- P_{BS} : Total power that a base station can send.
- v_s : Activity coefficient of a session.
- N_o : Spectral density of the noise.
- $a_{s,0}$: The attenuation factor between the mobile station s and the involved base station.
- $a_{s,i}$: The attenuation factor between the mobile station s and a near base station i.
- w : The orthogonality factor. The code are considered orthogonal to differentiate the users.
- $Ring(BS_0)$: The set of base stations in the neighborhood of the BS_0 that can serve the mobile station. As each cell has a hexagonal shape then there are six other base stations in the neighborhood of each base station.

Capacity formulas in the uplink direction

The assumptions of the model are:

- Interferences due to the cellular structure are taken into account, including both inter and intra-cell interferences. Assume a cellular system with users sharing the same radio bandwidth and using the same base station in each cell, each base station receives interferences from mobiles in the home cell called intra-cell interferences, and from mobiles or base stations located in the neighboring cells called inter-cell interferences. The intensity of the traffic should be homogeneous.
- We assume perfect power control.
- The power allocated to access channels is not taken into account.
- Cells are not sectorized, it means that there is only one sector per cell.

Starting from (4.1) and the definition of ratio of signal-to-noise, the following formula for the uplink direction is proposed by [5]:

$$\left(\frac{E_b}{N_t}\right)^{\text{UL}}_{s,r} = \frac{W}{r_t} \times \frac{P_s^{\text{UL}}}{((1+\lambda)P_T^{\text{UL}} - v_s P_s^{\text{UL}}) + WN_0}, \tag{4.5}$$

where:

- $P_T^{\text{UL}} - v_s P_s^{\text{UL}}$ is the intra-cell interferences caused by other sessions,
- λP_T^{UL} is the strength of inter-cell interferences.

Referring to the $\gamma_{s,r}^{\text{UL}}$ explained in the previous part, the value of $(\frac{E_b}{N_t})^{\text{UL}}_{s,r}$ should reach $\gamma_{s,r}^{\text{UL}}$ defined as follows:

$$\gamma_{s,r}^{\text{UL}} = \frac{W}{r_t} \times \frac{P_s^{\text{UL}}}{((1+\lambda)P_T^{\text{UL}} - v_s P_s^{\text{UL}}) + WN_0}. \tag{4.6}$$

From (4.6) we can extract P_s^{UL} as follows:

$$\gamma_{s,r}^{\text{UL}} r_t (1+\lambda) P_T^{\text{UL}} + \gamma_{s,r}^{\text{UL}} r_t W N_0 = P_s^{\text{UL}} (W - \gamma_{s,r}^{\text{UL}} r_t v_s),$$

$$P_s^{\text{UL}} = \gamma_{s,r}^{\text{UL}} \frac{(1+\lambda)P_T^{\text{UL}} + N_0}{\frac{1}{r_t} - \frac{\gamma_{s,r}^{\text{UL}} v_s}{W}}. \tag{4.7}$$

Recalling that $P_T^{\text{UL}} = \sum_s v_s P_s^{\text{UL}}$, we can replace P_s^{UL} by (4.7) and therefore we will have:

$$P_T^{\text{UL}} = \frac{\sum_s \frac{\gamma_{s,r}^{\text{UL}} v_s}{\frac{1}{r_t} + \frac{\gamma_{s,r}^{\text{UL}} v_s}{W}}}{1 - \frac{1+\lambda}{W} \sum_s \frac{\gamma_{s,r}^{\text{UL}} v_s}{\frac{1}{r_t} + \frac{\gamma_{s,r}^{\text{UL}} v_s}{W}}} \times N_0,$$

where we have defined the load factor, η_{UL}, as

$$\eta_{\text{UL}} = \frac{1+\lambda}{W} \sum_s \frac{\gamma_{s,r}^{\text{UL}} v_s}{\frac{1}{r_t} + \frac{\gamma_{s,r}^{\text{UL}} v_s}{W}}.$$

When η_{UL} becomes close to 1, the P_T^{UL} approaches to infinity therefore η_{UL} should be always less than 1.
The load factor plays a significant role in the capacity formula. We check that the sum of radio capacity of the sessions which are being served by a base station do not exceed the load factor. For this reason the authors of [5] suggest a limit up to 60 % of the capacity load on the WCDMA for UMTS. Using this limit on the load in the uplink direction, LF_{UL}, we satisfy:

$$\frac{1+\lambda}{W} \sum_s \frac{\gamma_{s,r}^{\text{UL}} v_s}{\frac{1}{r_t} + \frac{\gamma_{s,r}^{\text{UL}} v_s}{W}} < LF_{\text{UL}}. \tag{4.8}$$

According to (4.8) we have :

$$LF_{\text{UL}} \times \frac{W}{1+\lambda} > \sum_s \frac{\gamma_{s,r}^{\text{UL}} v_s}{\frac{1}{r_t} + \frac{\gamma_{s,r}^{\text{UL}} v_s}{W}}, \tag{4.9}$$

where $LF_{\text{UL}} \times \frac{W}{1+\lambda}$ defines the maximum capacity on uplink direction and $\frac{\gamma_{s,r}^{\text{UL}} v_s}{\frac{1}{r_t} + \frac{\gamma_{s,r}^{\text{UL}} v_s}{W}}$ is the contribution of each session on the capacity.

In the soft handoff case the minimum value of a signal, in order to be interpreted in the base station, reduces. This reduction will be called soft handoff profit and we denote this profit by SHG^{UL} for the uplink direction. The contribution of each session, taking in account the soft handoff profit, is equal to:

$$\frac{SHG^{\text{UL}}\gamma^{\text{UL}}_{s,r}v_s}{\frac{1}{r_t}+\frac{SHG^{\text{UL}}\gamma^{\text{UL}}_{s,r}v_s}{W}}. \tag{4.10}$$

Capacity formulas in downlink direction

The assumptions of the model for the downlink direction are:

- We assume perfect power control.
- Both inter-cell and intra-cell are taken into account.
- The traffic distribution is homogeneous.
- The attenuation is based on double reception of the signal, one is the signal which pass through the direct path from the sender to receiver and the other is the ground reflection.
- Cells are not sectorized.

In the downlink the inter-cell and intra-cell interferences are defined by $\frac{P_T^{\text{DL}}-v_sP_s^{\text{DL}}}{a_{s,0}}$ and $\sum\limits_{j\in Ring(\text{BS}_0)}\frac{P_T^{\text{DL}}}{a_{s,j}}$, therefore as for the uplink direction according to the formula (4.1) and the definition of ratio of signal-to-noise we have:

$$(\frac{E_b}{N_t})^{\text{DL}}_{s,r}=\frac{W}{r_t}\times\frac{\frac{P_s^{\text{DL}}}{a_{s,o}}\times\frac{1}{r_t}}{w\times\frac{P_T^{\text{UL}}-v_sP_s^{\text{DL}}}{a_{s,0}}+\sum\limits_{j\in Ring(\text{BS}_0)}\frac{P_T^{\text{DL}}}{a_{s,j}}+WN_0}. \tag{4.11}$$

As explained before a signal sent by a session s to a base station can reach the mobile station only if the $(\frac{E_b}{N_t})^{\text{DL}}_{s,r}$ of the signal received by mobile station is large enough. Referring to (4.1.4), let us assume that $\gamma^{\text{UL}}_{s,r}$ will be equal to this value. Therefore we have:

$$\gamma^{\text{UL}}_{s,r}=\frac{W}{r_t}\times\frac{\frac{P_s^{\text{DL}}}{a_{s,o}}\times\frac{1}{r_t}}{w\times\frac{P_T^{\text{UL}}-v_sP_s^{\text{DL}}}{a_{s,0}}+\sum\limits_{j\in Ring(\text{BS}_0)}\frac{P_T^{\text{DL}}}{a_{s,j}}+WN_0}. \tag{4.12}$$

As the total power sent by the base station to the attached mobile station (P_T^{DL}) is equal to the sum of the sent power by the base station to the accepted sessions plus the control and synchronization signal powers (P_{cont}) we have:

$$P_T^{\mathrm{DL}} = \sum_{s \in S} v_s P_s^{\mathrm{DL}} + P_{cont}. \tag{4.13}$$

We extract P_s^{DL} from the (4.12) formula and replace it in (4.13), then P_T^{DL} is:

$$P_T^{\mathrm{DL}} = \frac{P_{cont} + WN_0 \sum\limits_{s \in S} \frac{v_s \gamma_{s,r}^{\mathrm{DL}} a_{s,o}}{\frac{W}{r_t} + w v_s \gamma_{s,r}^{\mathrm{DL}}}}{1 - \sum\limits_{s \in S} \frac{\left(w + \sum\limits_{i \in Ring(\mathrm{BS}_0)} \frac{a_{s,o}}{a_{s,o}}\right) v_s \gamma_{s,r}^{\mathrm{DL}}}{\frac{W}{r_t} + w v_s \gamma_{s,r}^{\mathrm{DL}}}}. \tag{4.14}$$

As for the uplink direction, we can define the downlink load factor as:

$$\eta_{\mathrm{DL}} = \sum_{s \in S} \frac{\left(w + \sum\limits_{i \in Ring(\mathrm{BS}_0)} \frac{a_{s,o}}{a_{s,o}}\right) v_s \gamma_{s,r}^{\mathrm{DL}}}{\frac{W}{r_t} + w v_s \gamma_{s,r}^{\mathrm{DL}}}.$$

When η_{DL} gets close to 1, the system becomes instable and P_T^{DL} approaches infinity. However, the total power which can be sent by a base station in the downlink direction to different mobile stations (P_T^{DL}) should be smaller than the total power that a base station can send (P_{BS}). Therefore $P_T^{\mathrm{DL}} \leq P_{\mathrm{BS}}$ and from (4.14), we deduce:

$$\sum_{s \in S} \gamma_{s,r}^{\mathrm{DL}} v_s \times \frac{N_o a_{s,0} + \frac{P_{\mathrm{BS}}}{W}\left(\sum\limits_{i \in Ring(\mathrm{BS}_0)} \frac{a_{s,0}}{a_{s,i}} + w\right)}{\frac{\gamma_{s,r}^{\mathrm{DL}} v_s}{W} + \frac{1}{r_t}} \leq P_{\mathrm{BS}} - P_{cont}. \tag{4.15}$$

The inequality (4.15) is used to accept or reject a session on the radio link in the downlink direction, and the contribution of each session s in radio link capacity is:

$$\gamma_{s,r}^{\mathrm{DL}} v_s \times \frac{N_o a_{s,0} + \frac{P_{\mathrm{BS}}}{W}\left(\sum\limits_{i \in Ring(\mathrm{BS}_0)} \frac{a_{s,0}}{a_{s,i}} + w\right)}{\frac{\gamma_{s,r}^{\mathrm{DL}} v_s}{W} + \frac{1}{r_t}}. \tag{4.16}$$

As for the uplink direction, the minimum power of a signal, in order to be interpreted, would be reduced in soft handoff cases. We denote this reduction by SHG^{DL} as soft handoff profit factor in downlink. Therefore the contribution of a session in soft handoff case in radio link capacity turns to:

$$SHG^{\mathrm{DL}} \gamma_{s,r}^{\mathrm{DL}} v_s \times \frac{N_o a_{s,0} + \frac{P_{\mathrm{BS}}}{W}\left(\sum\limits_{i \in Ring(\mathrm{BS}_0)} \frac{a_{s,0}}{a_{s,i}} + w\right)}{\frac{\gamma_{s,r}^{\mathrm{DL}} v_s}{W} + \frac{1}{r_t}}. \tag{4.17}$$

More detailed explanations can be found in [5], [35] and [37].

4.1.5 Call Admission Control

Telecommunication networks aim to support integrated services over the low cost wireless services. For this reason resource sharing is considered as a major issue. Call Admission Control (CAC) strategies are used to limit the number of sessions for reducing congestion and call dropping, as well as satisfying the required Quality of Service (QoS). In other words, CAC algorithm ensures that the QoS of each connection can be maintained when a new connection is accepted. It decides whether a call can be accepted into the network based on the current traffic situation. We consider two types of call admission control for the real-time and non-real-time applications.

An end to end call admission algorithm is proposed in which the bandwidth of the flow is constant in the radio link and wired link parts of network during the life of a session. This algorithm makes its decision based on available resources on both wired links and radio links. If the needs of a session can be satisfied then the resources will be reserved on the link between the base station and the BSC and all the way on the path to the destination.

For each session s we consider a set of possible paths that can serve the session from the source to the destination which is an external node. The process of generating the set of possible paths is explained in Section 4.3. The available resources will be tested on both wired links and radio links in the two directions. If we consider ρ as the requested bandwidth for the session s, therefore for each possible path the residual bandwidth on each link of the path should be greater than ρ. We also make sure that for each direction (UL/DL) only and only one path among the paths in the set of possible paths will be selected. However, the chosen paths for uplink and downlink can be the same or may be different.

On the radio link the result of the test is positive if a base station can take care of the session without violating its available power and still being able to decode the signal.

Note that in the VBR case which are non-real-time applications, we have an upper and lower bound for the requested bandwidth on both radio links and wired links. Therefore, the residual bandwidth should be smaller or equal to upper bound and greater or equal to lower bound. These inequalities turn to equalities in CBR cases. All these test will be done for each session in radio links and wired links in both downlink and uplink directions.

Accepting and rejecting decision for a session

As mentioned before, for each session, the above tests should be done. A session may be accepted if the result of all the tests become positive, but it does not guarantee that the session will be accepted for sure. It depends at last on the GoS. The optimization procedure is global and it chooses certain sessions to be accepted. Therefore, if there exists another session with higher priority we may deny one with lower priority which has also positive results for the call admission control tests.

The number of rejected sessions should not go over a limit. This is the subject that we discuss under Grade of Service (GoS). The amount of GoS is defined for each (application, priority). The dimensioning procedure should also satisfy the quality of service of accepted sessions.

4.2 Traffic modeling

As 3G network promises to support various kinds of sevice classes, see Section 2.7, we are going to consider different types of applications, both real-time and non-real-time categories. We use a Constant Bit Rate (CBR) for real-time applications. It means that a real-time session has always the same constant bandwidth. Variable Bit Rate (VBR) is defined for non-real-time applications, with upper and lower bounds on the possible bandwidth.

In this traffic model, both interactive and non interactive applications are discussed. Recall that the interactive applications are the ones with both uplink and downlink directions since the source and destination interact simultaneously. Voice, videophone and web browsing are examples of this kind of application. On the other hand, mail and video streaming are non iterative since there is only in one direction at a time. Note that sessions corresponding to video streaming applications exist only in the downlink direction but mail sessions can be either in the uplink or in the downlink direction.

Another parameter considered in this traffic model is symmetric or asymmetric flows. A flow is called symmetric when it has the same bandwidth in both directions: voice and videophone flows have this characteristic. For web browsing sessions, however, the bandwidth of the flow on the uplink and downlink are not the same and are therefore asymmetric flows. As an example, the request for downloading a page which happens in uplink direction (from a mobile user to a server) needs less bandwidth and is much faster comparing to process of downloading the page that takes place in downlink direction (from server to mobile user). Table 4.2 describes the applications taken into account in our mathematical model.

Some new notations are used: the *High* and *Low* parameters define the level of requested QoS. This mainly concerns the bandwidth and the delay of each application. For an application such as video streaming in the *High* case we need to assign more bandwidth and less delay comparing to the same application in the *Low* case. Therefore, the

Application	**Category**	**Class**	**Bit Rate**	**Priority**	**Flow**
Voice	Real-time	Conversational	CBR	Gold/Silver	UL/DL Interactive Symmetric
Videophone High/Low	Real-time	Conversational	CBR.	Gold/Silver	UL/DL Interactive Symmetric
Video Streaming High/Low	Real-time	Streaming	CBR.	Gold/Silver	DL Non interactive
Web Browsing High/Low	Non-real-time	Interactive	VBR.	Gold/Silver	DL or UL Interactive Asymmetric
Mail High/Low	Non-real-time	background	VBR.	Gold/Silver	DL or UL Non Non interactive Symmetric

TAB. 4.2: Applications in the traffic model

network supports with a higher priority the sessions with a *High* specification as their requested QoS are more sensible in terms of bandwidth and delay.

Each session is assumed to have a priority, *Gold* or *Silver*. This has also an effect on the acceptance or rejection of a session. It means that the chance of rejecting of a session with Gold priority is less than the same type of session with Silver priority.

4.2.1 Generation of sessions

A session concerns a connection between a mobile user and an external node which is a fixed node in the external network. The number of requested sessions are known from the beginning of the simulation time, so the traffic can be placed on the dimensioning surface. The sessions of the same type and beginning period time, will be placed uniformly on the geographical surface. According to the position of each session we define a set of potential base stations for each session. This set is divided into two subsets. One which can serve the session only in non soft handoff and the other which can serve the session both in non soft handoff and soft handoff. The optimization procedure makes sure that if all constraints are satisfied one or more base stations can take care of a session from its set of potential base stations.

The corresponding BSC for each session is not predefined. For each individual session, we may have different possible paths through different BSC, see Section 4.3. The traffic load varies during different periods of a day. In one period we may have more voice sessions while in another, more web browsing sessions are requested. Some periods are very quiet (during the night) while other periods may be highly loaded (in the afternoon). That is why the planning time is divided into periods and each session covers a given number of periods of the planning periods. The duration of some sessions, such as mail, is short while others such as web browsing, can be longer. For example in Figure 4.3 the planning time is divided into seven periods and is equal to $P1 + P2 + P3 + P4 + P5 + P6 + P7$.

We assume that the beginning and the finish time of each session coincide with those of the periods, Figure 4.3. It means that the mathematical model reserves the resources

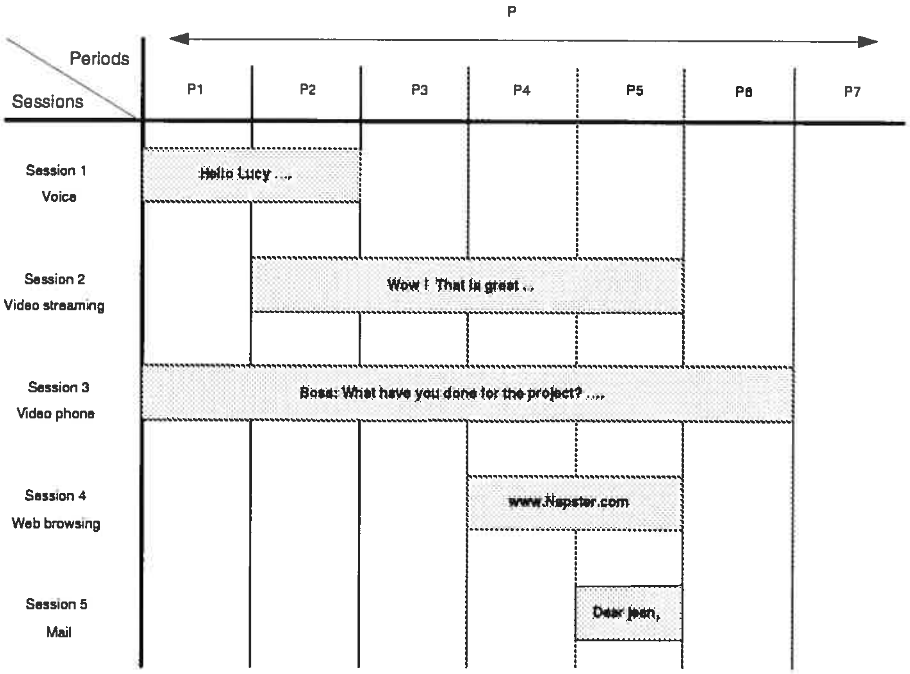


FIG. 4.3: Temporal sequencing

during the whole period if a session is accepted even if it has not started from the beginning of that period.

In the traffic model the number of sessions, their durations and their beginnings are assumed to be known. For instance, we know that the first session which is a voice needs to be served from the first period during 2 periods while the second session which is a video streaming starts at the second period and lasts for 4 periods. Here the anticipative call admission control test helps to do the network dimensioning task. If the reservation of resources can be done successfully in both wired and radio links during the life time of a session, the session is accepted.

Flow generation and their bandwidth for each session

As shown in Figure 4.4 different flows are engaged while a session is requested. The number of flows depend on the type of the session, interactive or non interactive. In

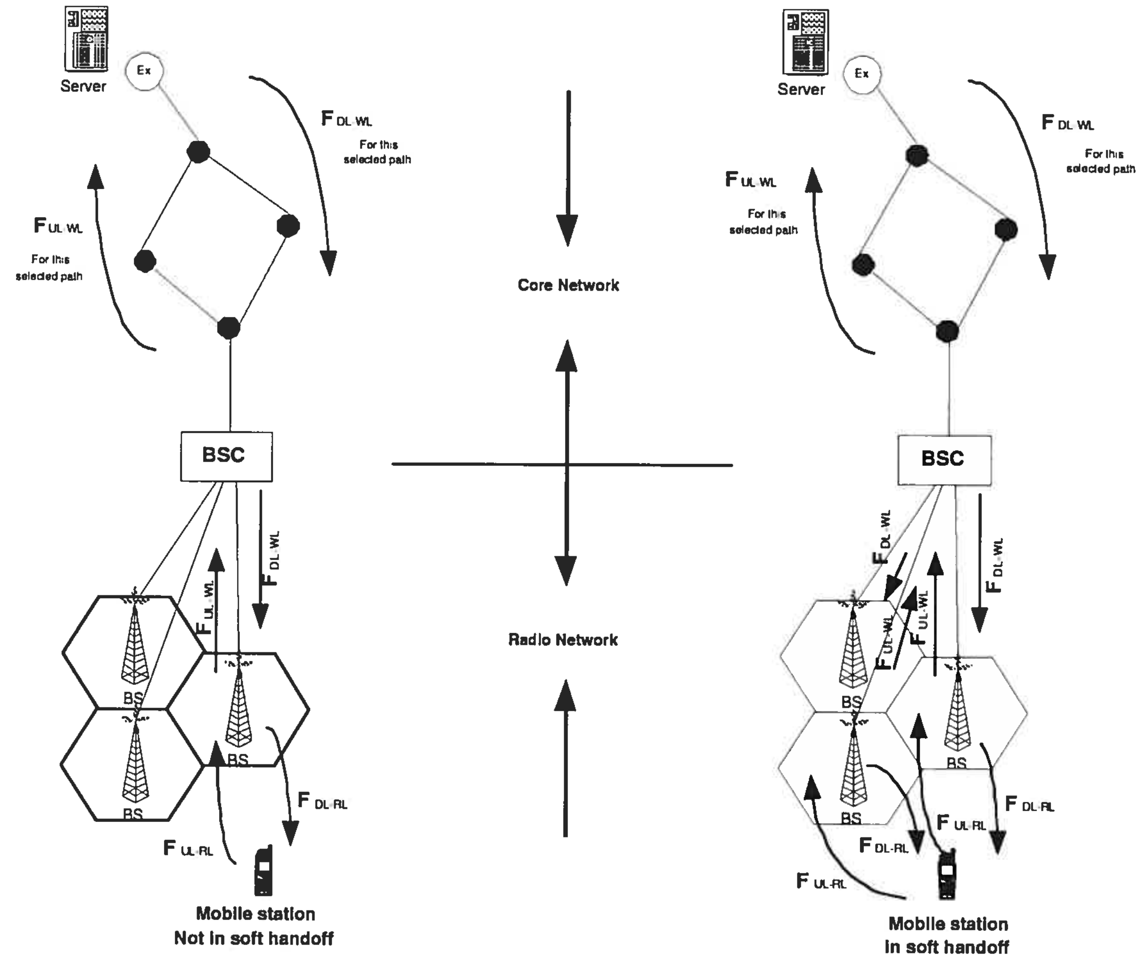


FIG. 4.4: Flows

general we can divide the flows into two classes:

➢ flow from mobile station to external node named uplink, f_{UL}.

➢ flow from external node to mobile station named downlink, f_{DL}.

Each interactive session, such as voice, would have both classes of flow but non interactive sessions, such as mail, would have either uplink or downlink classes of flows. Both of these classes are divided again into two different types of flow:

- uplink flow in wired links, $f_{\mathrm{UL\_WL}}$.
- downlink flow in wired links, $f_{\mathrm{DL\_WL}}$.
- uplink flow in radio links, $f_{\mathrm{UL\_RL}}$.
- downlink flow in radio links, $f_{\mathrm{DL\_RL}}$.

We now discuss the assigned bandwidth to each flow of a session. As mentioned before there are two different types of applications, CBR and VBR, based on the bit rate of the session during its life time. We discuss the bandwidth of flows in each type one after the other.

- CBR Applications:

  This is the simple case, since the bandwidth on the wired and radio links are always the same. For each CBR session there are the set of RABs with the same bandwidth but different Frame Error Rates (FER). The model will choose one of the sets, based on the priority of the session. The lower the FER, the better to be chosen for higher priority sessions. It has also the impact on the Quality of Service as it has an influence on the transmission delay on the radio links.

  It seems that this process is not that much complicated as the optimization model will choose the determined bandwidth with a proper FER.

- VBR Applications:

  In the Variable Bit Rate applications the stream is not fix, that is, sometimes the required bandwidth is low and other times it is high. There are also variable bandwidths on the wired and radio links. However, for each flow the maximum (upper bound) and the minimum (lower bound) of the bandwidth are fixed. Therefore the optimization model chooses a bandwidth between the lower and upper bounds on wired links and a RAB among the set of possible RABs for radio links.

In addition to this specification we propose some relations between the bandwidth on wired links and radio links.

On uplink direction, the bandwidth on the wired links $f_{\mathrm{UL\_WL}}$ are greater or equal to the bandwidth on the radio links $f_{\mathrm{UL\_RL}}$ means Data Rate$[f_{\mathrm{UL_WL}}] \geq$ Data Rate$[f_{\mathrm{UL_RL}}]$.

On downlink direction, the bandwidth on the radio links $f_{\mathrm{DL\_RL}}$ are greater or equal to the bandwidth on the wired links $f_{\mathrm{DL\_WL}}$ means Data Rate$[f_{\mathrm{DL_RL}}] \geq$ Data Rate$[f_{\mathrm{DL_WL}}]$.

These rules prevent the accumulation of the packets at the base stations level. The second rule mentions that in the downlink, $f_{\mathrm{DL\_WL}}$ should be smaller than $f_{\mathrm{DL\_RL}}$. In order to prevent choosing a very small value for $f_{\mathrm{DL\_WL}}$ we propose a set of discrete values for each RAB r in radio links named $\mathcal{R}(r_t)$. It means that for each selected RAB r for the flow on radio link on downlink direction then the bandwidth on wired link should not be less than a certain value. Table 4.3 shows the value of $\mathcal{R}(r_t)$ for each debit of RAB r. For simplicity we show these rules as the following inequalities:

Radio		Core
RAB throughput	FER	$\mathcal{R}(r_t)$
9.6 kbps	1% , 2%	-
19.2 kbps	5% , 10%	10 kbps
38.4 kbps	1% , 5% , 10%	20 kbps
76.8	1% , 5% , 10%	40 kpbs
153.6 kbps	1% , 5% , 10%	80 kpbs
307.2 kbps	5% , 10%	160 kpbs

TAB. 4.3: Table of values for $\mathcal{R}(r_t)$.

$$\text{Data Rate}[f_{\mathrm{UL\_WL}}] \geq \text{Data Rate}[f_{\mathrm{UL\_RL}}] \tag{4.18}$$

$$\text{Data Rate}[f_{\mathrm{DL\_RL}}] \geq \text{Data Rate}[f_{\mathrm{DL\_WL}}] \tag{4.19}$$

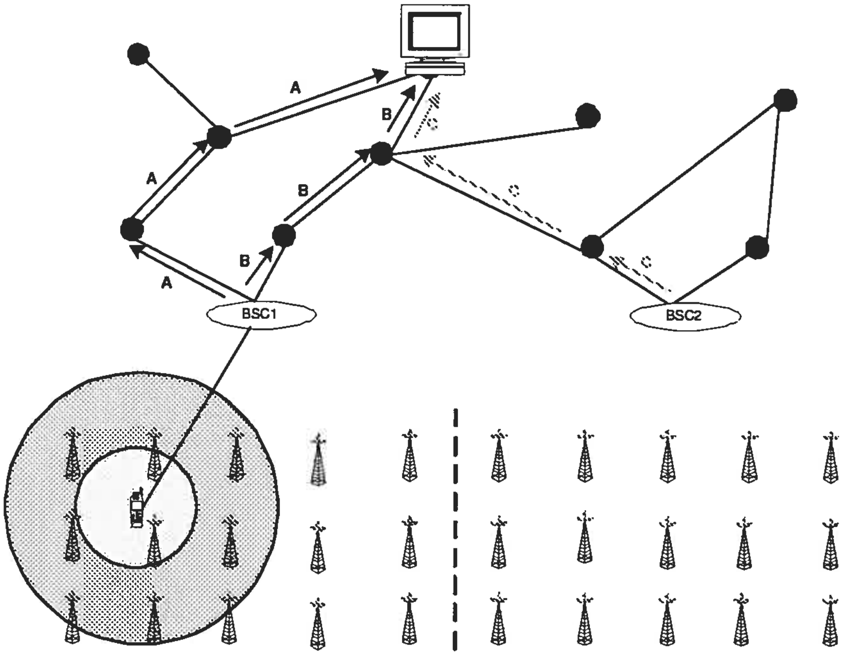
$$\text{Data Rate}[f_{\text{DL\_RL}}] \geq \mathcal{R}(r_t) \qquad when \qquad \text{Data Rate}[f_{\text{DL\_RL}}] = r_t \tag{4.20}$$

4.3 Routing

First of all we should clarify the concept of a path or a route in this mathematical model. A route or a path (we may use one of these words in all the thesis) is a set of wired links between a BSC and an external node. All the external nodes are mentioned in the network architecture file as well as the possible paths from each BSC to each external node. On the BSC level, there is a routing table which gives us the possible routes to the destination of a session.

The next step is the selection of the list of potential paths among the existing paths which are destined to the selected external node. For this reason, we have to find out which BSC or BSCs can support the session. Each session has a mother-BSC. The mother-BSC for each session is the BSC which covers its geographical position. Therefore all the paths in the mother-BSC routing table which go to the destination of that session can be taken as potential paths. Now we check if there is another BSC which can serve the session rather than its mother-BSC. For this reason, we first find the potential base stations who can give service to this session in both soft handoff and non-soft handoff. This is based on the location of the session and the distance between the session and the base stations. At this point we check the responsible BSC for each potential base station. If all the base stations are connected to the mother-BSC of that session (for instance BSC=1), then it means that only the paths which pass through BSC=1 to the destination can be in the set of potential paths, see Figure 4.5. On the other hand, if there are potential base stations which are connected to a different BSC other than the mother-BSC then the paths which go to the destination and pass that BSC, will be added to the list of possible paths for that session, see Figure 4.6.

There is a sorting procedure that sorts the set of routes. If the selected routes are with different lengths we sort the set from the shortest to the longest. Shortest means the path with less number of links. This means that the shortest path is always at the head of the set. Note that the first path will be always the one which passes the mother-BSC.



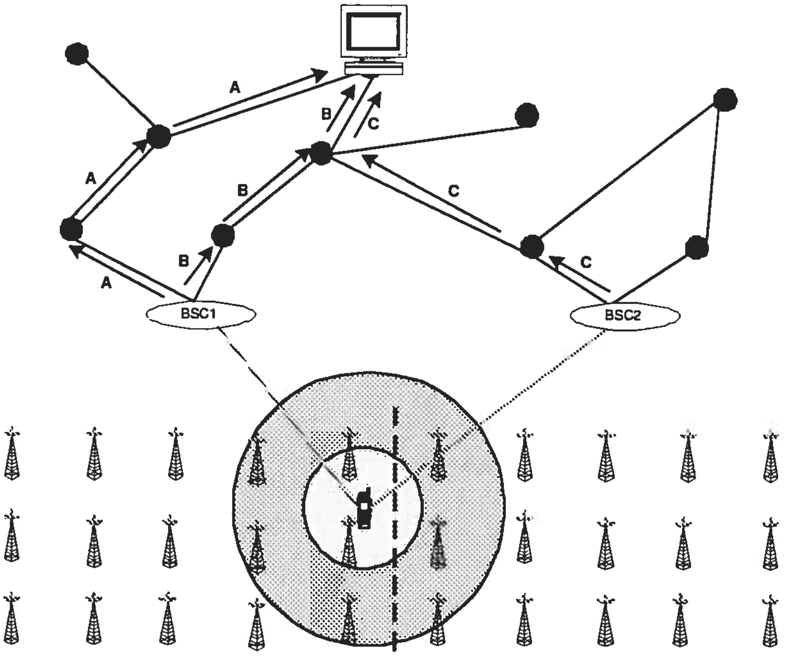
Possible paths { A, B}
Mother BSC : BSC 1

FIG. 4.5: Selection of paths I

Then, the next step concerns paths with the same lengths. The position of these routes in the set will be chosen randomly.

Consequently, for each individual session we determine a set of possible routes, among all the routes, concerning the destination and the position of the session. With this routing strategy first the unrelated routes can be avoided to be taken in account in the mathematical model. Second, considering a set of possible paths that may be able to fulfill the QoS constraints is much more realistic than considering just the shortest path or a random path.

The hop by hop IP packet mechanism, for real-time applications, which is currently used on the Internet is simple and scalable. It leads to the success of Internet but is not enough to provide QoS. Therefore, it is not suitable for the services in the new generation of networks. While the proposed model in this thesis fits in 3G and 4G networks. In this study once a path is selected to serve a session it stays the same during the



Possible paths { A, B, C}
Mother BSC : BSC 1

FIG. 4.6: Selection of paths II

life of the session. Our reasoning is based on satisfying the demands on 3G networks. Considering the services which asks for high bandwidth and QoS checking requirements and RSVP informations, hop by hop processing, makes each router a bottleneck and violates the QoS. That is the reason that we have mainly focused on the routing on the same path from the source to the destination.

4.4 Delay

As mentioned in Table 2.2 each type of applications has different sensibilities to delay. The delay is critical for the real-time applications: they provide an action or an answer to an external event in a time with predictable manner. It means that the packets should get to the destination with a short delay of time after they leave the source. There is no retransmission in the case of error and a packet with error is discarded.

On the other hand, for the non-real-time applications the integrity of data is much more important than the arrival time. Actually in the mail, for example, the sender does not expect any answer back sooner than a certain time, but he does not accept any alter in his mail. Therefore, these kinds of applications are not delay sensitive but error sensitive. In addition the rejected or errored packets can be retransmitted over the network. Note that the retransmission delay on the wired links are negligible and we have not considered them in the model.

The delay is going to be considered for the radio links and wired links, and each of them has its own specifications. On both parts, the delay on uplink (D_s^{UL}) and downlink (D_s^{DL}) directions would be studied. It means that the delay in the uplink is equal to the sum of delays both in wired links and radio links in the uplink direction. The same concept applies for the downlink as well. The maximum transmission delay is considered and therefore the minimum bandwidths are imposed for different flows. As the bandwidth is fixed for the CBR applications, delay constraint is taken in account only for the VBR applications in which the procedure chooses a bandwidth for the flow of each session.

4.4.1 Delay on wired link part

Delay on the wired link is the sum of transmission delay, propagation delay, queuing delay corresponding to the WFQ policy and delay on the routers (treatment delay) in both directions.

We can summarize this explanation in the two following formulas:

$$D_{f^{\text{UL\_WL}}} = D^{trans}_{f^{\text{UL\_WL}}} + D^{propag}_{f^{\text{UL\_WL}}} + D^{treat}_{f^{\text{UL\_WL}}} + D^{queuing}_{f^{\text{UL\_WL}}},$$

$$D_{f^{\text{DL\_WL}}} = D^{trans}_{f^{\text{DL\_WL}}} + D^{propag}_{f^{\text{DL\_WL}}} + D^{treat}_{f^{\text{DL\_WL}}} + D^{queuing}_{f^{\text{DL\_WL}}}. \tag{4.21}$$

- Transmission delay:

  This delay concerns the time that takes a packet to go to a defined destination from a defined source. The transmission delay between two nodes is generally calculated by the following formulas in downlink and uplink:

$$D^{trans}_{f_{\text{DL\_WL}}} = \frac{\text{size of DL packet}}{\text{bandwidth}},$$

$$D^{trans}_{f_{\text{UL\_WL}}} = \frac{\text{size of UL packet}}{\text{bandwidth}},$$

where we consider a fixed size of packet for each application.

Since, in downlink the minimum data rate on the cabled links is the $\mathcal{R}(r_t)$ and a packet may pass through several intermediate nodes to reach to the destined node, the transmission delay formula in downlink direction can be changed to:

$$D^{max-trans,r_t}_{f_{\text{DL\_WL}}} = m_{f_{\text{DL}}} \times \frac{\text{size of DL packet}}{\mathcal{R}(r_t^{\text{DL}})}, \tag{4.22}$$

where $m_{f_{\text{DL}}}$ is the number of links on the downlink selected path between the source and destination nodes. While in the uplink direction the minimum bandwidth on the wired link has the value of the bandwidth of r_t^{UL} and if we consider $m_{f_{\text{UL}}}$ as the number of links of the selected path in this direction we would have:

$$D^{max-trans}_{f_{\text{DL\_WL}}} = m_{f_{\text{UL}}} \times \frac{\text{size of UL packet}}{r_t^{\text{UL}}}. \tag{4.23}$$

- Propagation delay:

This delay concerns the propagation delay of bits on the links. It depends on the distance between each node and the speed of light:

$$D^{propag}_{f_{\text{WL}}} = \frac{\text{distance between two nodes}}{\text{speed of light}}.$$

Like the previous section if we consider the number of links between each source and destination nodes in downlink as $m_{f_{\text{DL}}}$, and in uplink as $m_{f_{\text{UL}}}$ with an average propagation time for each link, we would have:

$$D^{propag}_{f_{\text{DL\_WL}}} = m_{f_{\text{DL}}} \times \text{propagation time/link}, \tag{4.24}$$

$$D^{propag}_{f_{\text{UL\_WL}}} = m_{f_{\text{UL}}} \times \text{propagation time/link}. \tag{4.25}$$

- Treatment delay:

  When a packet leaves a source node toward a destination node it passes through intermediate nodes which we call routers. The error detection, correction and routing procedure to the next node would take place on the level of each router. Obviously, this would cause the treatment delay. Since this delay would not occur in the source node, we can again use the $m_{f_{\mathrm{DL}}}$ as the number of intermediate nodes in downlink and $m_{f_{\mathrm{UL}}}$ as the number of intermediate nodes in uplink direction.

$$D^{treat}_{f_{\mathrm{DL\_WL}}} = m_{f_{\mathrm{DL}}} \times \text{treatment time/router}, \tag{4.26}$$

$$D^{treat}_{f_{\mathrm{UL\_WL}}} = m_{f_{\mathrm{UL}}} \times \text{treatment time/router}. \tag{4.27}$$

- Queuing delay:

  Packet-by-packet Generalized Processor Sharing (PGPS) is the name given to the method for computing the queuing delay bound in integrated services under WFQ scheduling policy. The packets are queued in each router on the path which goes to the destination. Thanks to the [38] with the PGPS when the RPPS (Rate Proportional Processor Sharing) discipline is used, there exists a maximum limit on the queuing delay from the source to the destination where we put a leaky bucket (Appendix A) on the entrance node to wired link part for each session. Leaky bucket allows the separation of delays to delay in leaky bucket and delay in the network. [38], [39] have concentrated on providing performance guarantees on throughput and worst case packet delay in single node case and multiple nodes case. This delay can be calculated by $D^* = \frac{\sigma+(k-1)L}{\rho}$ where k is the number of nodes on the path excluding the destination node ($(k-1)$ is equal to the number of links on the path plus the link between base station and its corresponding BSC), L is the size of the packet, ρ is the data rate of the source and σ is the maximal number of packets which can be in the leaky bucket. Now, we can adopt the formulas below for uplink and downlink directions:

$$D^{queuing}_{f_{\mathrm{DL\_WL}}} = \frac{\sigma + m_{f_{\mathrm{DL}}} \times \text{size of packet}}{\rho}, \tag{4.28}$$

$$D^{queuing}_{f_{\mathrm{UL\_WL}}} = \frac{\sigma{+}m_{f_{\mathrm{UL}}} \times \text{size of packet}}{\rho}. \tag{4.29}$$

The mobile user does not want to wait. When it asks for a session, that session can be accepted or rejected (the blocking rate in GoS). On the other hand, in the uplink direction, in order to prevent the accumulation on the base stations, the bandwidths in wired links are greater than the bandwidths in radio links, formula (4.18). We also reserve the maximum bandwidth on the wired link as the bandwidth of the RAB reaches the base station has the maximum amount. It is the leaky bucket, in which the token rate is equal to bandwidth of the RAB and the bucket size is zero. Therefore, we conclude that there is no need to have the delay on uplink direction and we would concentrate just on the downlink direction.

While, in downlink direction where the leaky buckets are placed on the access nodes, source sends the IP packets to the bucket. Therefore the queuing delay will not be equal to zero and the maximum limit of delay will be used.

$$D^{queuing}_{f_{\text{DL\_WL}}} \leq D^*,$$

Replacing the appropriate formulas in (4.21) leads us to:

$$D_{f_{\text{DL\_WL}}} \leq D^{trans,rt}_{f_{\text{DL\_WL}}} + D^{propag}_{f_{\text{DL\_WL}}} + D^{treat}_{f_{\text{DL\_WL}}} + \frac{\sigma + m_{fDL} \times \text{size of packet}}{\rho}. \tag{4.30}$$

4.4.2 Delay on radio link part

The duration of a frame in CDMA2000 is equal to 20 ms (0.2 s) therefore the number of bits in a frame is: 0.2 × data rate of RAB (since the data rate of RAB is bit per second). If we consider the $N_{r_t^{\text{UL}}}$ and $N_{r_t^{\text{DL}}}$ as the number of frames per packet with the data rate of r_t^{UL} and r_t^{DL} on uplink and downlink directions we have:

$$N_{r_t^{\text{UL}}} = \frac{8 \times \text{size of packet (byte)}}{\text{frame duration} \times r_t^{\text{UL}}}, \tag{4.31}$$

$$N_{r_t^{\text{DL}}} = \frac{8 \times \text{size of packet (byte)}}{\text{frame duration} \times r_t^{\text{DL}}}. \tag{4.32}$$

When a frame is lost or there is an error in that frame we use the Selective Repeat Protocol (SRP), see Appendix B, as the retransmission of the frame on the radio link.

The probability of an error on the frame depends on Frame Error Rate (FER) and the selected RAB. The transmission delay of a frame when at least three retransmissions are needed is given on Table 4.4. All the calculations are available in Appendix B:

FER	$D_{r_{\text{FER}}}^{radio-transmission}$
0.5%	101.01 ms
1%	102 ms
2%	104.11 ms
5%	110.68 ms
10%	122.78 ms

TAB. 4.4: Radio transmission delay of a frame

Since the transmission delay of an IP packet is equal to the sum of the transmission delay of all its frames we conclude:

$$D_{f^{\text{UL\_RL}}}^{r} = N_{r_t^{\text{UL}}} \times D_{r_{\text{FER}}}^{radio-transmission}, \tag{4.33}$$

$$D_{f^{\text{DL\_RL}}}^{r} = N_{r_t^{\text{DL}}} \times D_{r_{\text{FER}}}^{radio-transmission}. \tag{4.34}$$

4.5 Dimensioning strategy

For each simulation the number of sessions, potential base stations, BSCs and network topology are known from the beginning. The distribution of the sessions, BSCs and potential base stations are discussed in Sections 4.2.1 and 4.1.2. The proposed strategy does not let a base station serves a session which is situated far from it. It also does not accept a base station in soft handoff when it is placed so close to the session. The optimization procedure is global and considers all the periods at the same time for accepting or rejecting a session. It makes sure that even for the loaded traffics we will have a solution. Each serving base station that even take charge of one session will be added to the list of activated base stations. The optimization procedure also looks for

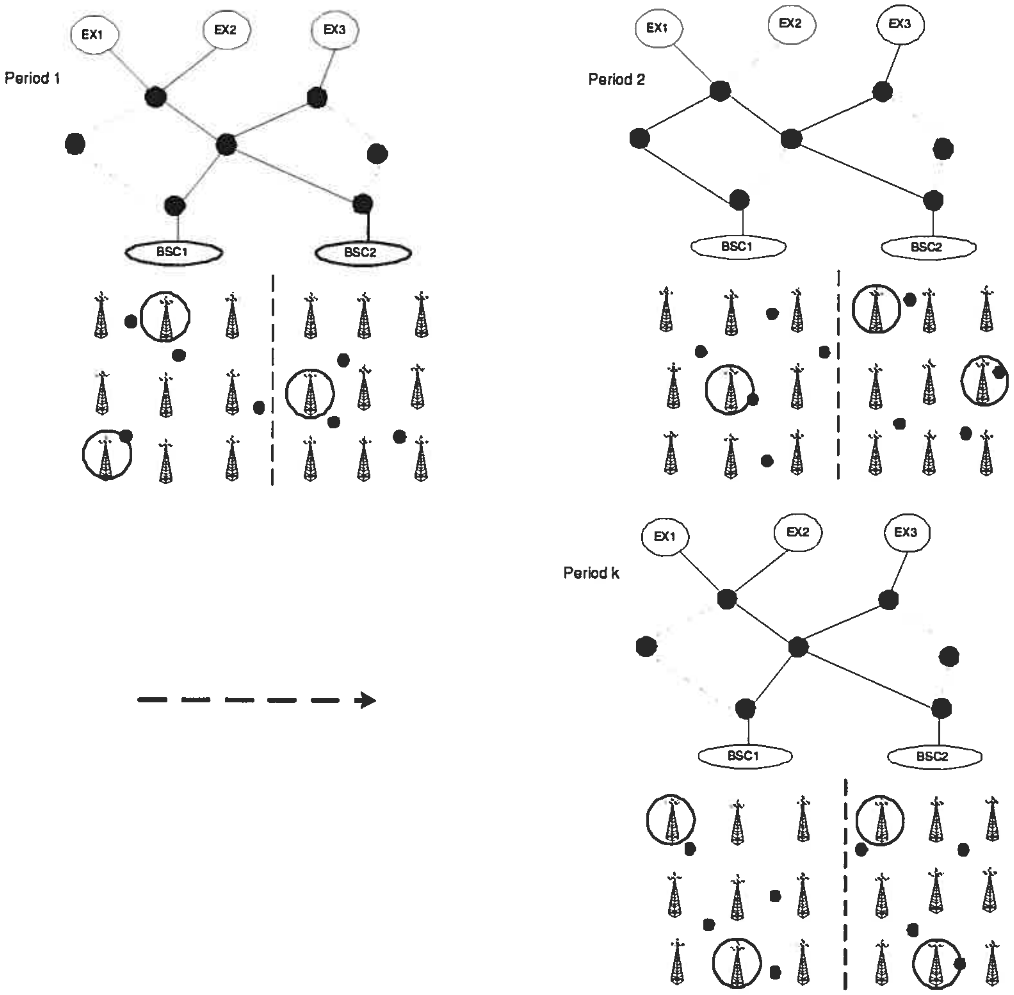


FIG. 4.7: Dimensioning procedure

the best path among the set of possible paths, Section 4.3, in order to minimize the capacity on the wired links. The selected paths for the uplink and downlink direction of a session are not necessarily the same. However, in each direction the selected path stays the same, during the life time of a session. The dimensioning strategy is illustrated in Figures 4.7 and 4.8.

The objective of the dimensioning includes both the wired links and radio links parts:

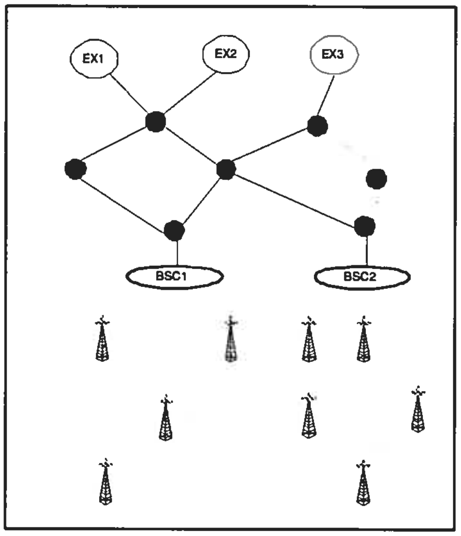


FIG. 4.8: The result of dimensioning

➤ Wired links:

We try to minimize the used capacity in the wired links by choosing the best path among the set of possible paths. Therefore, at the end of the optimization procedure the maximum capacity used in each wired link to serve the mobile stations will be its optimal value. Talking about wired links, we mean the links in core network and the links between the base stations and their corresponding BSC in radio network. Depending on the chosen path to the destination for each session and the type of that session the model finds out the capacity on the links. The network architecture is known from the beginning of the simulation.

➤ Radio links:

We try to minimize the number of activated base stations during the planning time in the optimization process. This is a critical issue as the base stations are so expensive to be used without enough studies before their installation. At the end of the optimization the number of activated base stations and their geographical positions will be revealed. The procedure chooses the ones which can serve the mobile stations and respect the Quality of Service and Grade of Service. Note that, the maximum number of potential base stations and their geographical positions are known form the beginning of the simulation. Once a base station serves one session it will be added to the activated base stations list. For each session, the serving base station will be determined by the capacity formulas in radio links.

Chapter 5

Mathematical Model

This chapter is devoted to a detailed description of the mathematical model that we have developed for the dimensioning of 3G networks. We started from the mathematical model proposed in the M.Sc. of C. Voisin published in [1]. We then made several improvements to this model and generalized it for multi-routing strategies. In addition to the generalization, other modifications have been made in order to improve the original model in [1]. The improvements help to remove some assumptions made in [1] in order to make the model much more realistic.

- Soft handoff between more than two base stations.
- The potential base stations which can serve a session may be connected to different base station controllers.
- Quality of service constraints: Selection of RAB in radio link.

In the generalized and improved model developed in this thesis we have tried to reduce the addition of new binary variables as much as possible and keep the model linear. We also attempted as much as possible to keep the model at the same level of complexity even if we are considering multi-routing. The new constraints and the constraints which have been modified comparing to [1] are identified by "*" in Section 5.6.

5.1 Notations

5.1.1 General notations

s	a session,
S	a set of sessions,
f	a flow,
F	a set of flows,
$F^{\text{UL\_WL}}$	a set of flows on the wired links in uplink direction,
$F^{\text{DL\_WL}}$	a set of flows on the wired links in downlink direction,
$F^{\text{UL\_RL}}$	a set of flows on the radio links in uplink direction,
$F^{\text{DL\_RL}}$	a set of flows on the radio links in downlink direction,
P	set of priority for the users,
A	set of different applications,
S^p	set of sessions with priority p,
S^a	set of sessions of application a,
S^{CBR}	set of sessions of CBR applications,
S^{VBR}	set of sessions of VBR applications,
H	set of periods of time,
h	a period.

5.1.2 Wired link parameters

m number of wired links in the core network,

m_v number of links in the path v,

d_ℓ length of link ℓ in the core network,

$d_{\ell_{\mathrm{BS}i}}$ length of link ℓ which connects the base station i to its BSC,

K number of existing base station controller (BSC),

U maximum number of base stations participating in the soft handoff.

5.1.3 Routing parameters

v potential path in core network i.e. a path between a BSC and a destination node defined as a set of links,

V_f^k the set of potential paths for flow f connecting the external node of the session to the BSC_k.

This set can be different for uplink and downlink. We also define:

$$V_f = \bigcup_{k=1}^{K} V_f^k. \tag{5.1}$$

$$a_{hs} = \begin{cases} 1 & \text{if session } s \text{ exists during period } h. \\ 0 & \text{otherwise.} \end{cases}$$

$$m_{vl} = \begin{cases} 1 & \text{if link } l \text{ belongs to path } v \\ 0 & \text{otherwise.} \end{cases}$$

$$b_{ihf} = \begin{cases} 1 & \text{if } BS_i \text{ belongs to the set of potential base stations that can carry flow} \\ & f \text{ during period } h \\ 0 & \text{otherwise.} \end{cases}$$

5.1.4 Session parameters

F_s set of all the four flows described below,

$f_s^{\text{UL\_WL}}$ flow on the wired links in the uplink direction for the session s,

$f_s^{\text{DL\_WL}}$ flow on the wired links in the downlink direction for the session s,

$f_s^{\text{UL\_RL}}$ flow on the radio links in the uplink direction for the session s,

$f_s^{\text{DL\_RL}}$ flow on the radio links in the downlink direction for the session s,

$\overline{\rho}_f$ maximum bandwidth for flow f on the wired links when session is VBR,

$\underline{\rho}_f$ minimum bandwidth for flow f on the wired links when session is VBR,

$\overline{t}_f$ maximum bandwidth for flow f on the radio links when session is VBR,

$\underline{t}_f$ minimum bandwidth for flow f on the radio links when session is VBR,

ρ_f declared bandwidth for the flow f on the wired links when session is CBR,

t_f declared bandwidth for the flow f on the radio links when session is CBR,

$D_s^{\text{DL}-MAX}$ maximum end to end transfer delay in downlink of a packet with the flow $f_s^{\text{DL\_WL}}$,

$D_{f_s^{\text{DL\_WL}}}^{max-transe,rt}$ maximum transmission delay in downlink of a packet with the flow $f_s^{\text{DL\_WL}}$ when the RAB r is selected,

$D_{f_s^{\text{DL\_WL}}}^{max-propag}$ maximum propagation delay for a packet with flow $f_s^{\text{DL\_WL}}$,

$D_{f_s^{\text{DL\_WL}}}^{max-trait}$ maximum treatment delay for a packet with flow $f_s^{\text{DL\_WL}}$,

L_f^{max} maximum size of a packet with the flow f,

$T_{a,p}^{\text{UL}}$ set of possible RABs for a session of application a with priority p in uplink direction in the radio network,

$T_{a,p}^{\text{DL}}$ set of possible RABs for a session of application a with priority p in downlink direction in the radio network,

T_s^{UL} set of possible RABs for session s in uplink direction ($T_s^{\text{UL}} \subseteq T_{a_s,p_s}^{\text{UL}}$),

T_s^{DL} set of possible RABs for session s in downlink direction ($T_s^{\text{DL}} \subseteq T_{a_s,p_s}^{\text{DL}}$),

L_s^{BS} set of potential base stations which can give service to the session s in soft handoff and non soft handoff,

$L_s^{\text{BS\_SH}}$ set of the potential base stations which can give service to session s on soft handoff ($L_s^{\text{BS\_SH}} \subseteq L_s^{\text{BS}}$),

N_{BS_k} set of potential base stations connected to the BSC_k,

K_s set of BSCs which their corresponding BSs can serve the session in soft handoff or non soft handoff,

K_s^{SH} set of BSCs which their corresponding BSs can serve the session just in soft handoff.

Quality and grade of service parameters

$\text{Br}_{a,p}$ blocking rate for flows of application a with priority p,

$Q_{r,a,p}^{\text{UL}}$ Quality of Service coefficient for sessions of application a and priority p and served with RAB r on the uplink,

$Q_{r,a,p}^{\text{DL}}$ Quality of Service coefficient for sessions of application a and priority p and served with RAB r on the downlink.

5.2 Definition of the variables

- x_f^v : bandwidth of flow f on path v .

$x_{f_s^{\text{UL\_WL}}}^v$ = bandwidth of flow for the uplink of wired links on the path v.

$x_{f_s^{\text{DL\_WL}}}^v$ = bandwidth of flow for the downlink of wired links on the path v.

- x_f : bandwidth of flow f in Radio Links.

$x_{f_s^{\text{UL\_RL}}}$ = bandwidth of flow for the uplink of Radio Link network.

$x_{f_s^{\text{DL\_RL}}}$ = bandwidth of flow for the downlink of Radio Link network.

- x_f^i :bandwidth of flow f ($\in F_s^{\text{WL}}$) on the link between the base-station BS_i and its corresponding BSC in the radio network.

- $\alpha_{irf} = \begin{cases} 1 & \text{if the base station } i \text{ and the RAB } r \text{ are selected for the flow} \\ & f \text{ of session } s. \\ 0 & \text{otherwise.} \end{cases}$

- $y_i^{\mathrm{BS}} = \begin{cases} 1 & \text{if the base station } i \text{ has been selected at least once to serve a session.} \\ 0 & \text{otherwise.} \end{cases}$

- $\beta_{vf_s^{\mathrm{UL\text{-}WL}}} = \begin{cases} 1 & \text{if path } v \text{ is chosen for flow } f_s^{\mathrm{UL\text{-}WL}}. \\ 0 & \text{otherwise.} \end{cases}$

- $\beta_{vf_s^{\mathrm{DL\text{-}WL}}} = \begin{cases} 1 & \text{if path } v \text{ is chosen for flow } f_s^{\mathrm{DL\text{-}WL}}. \\ 0 & \text{otherwise.} \end{cases}$

- $y_s = \begin{cases} 1 & \text{if all flows } f \in F_s \text{ are satisfied} \\ 0 & \text{otherwise.} \end{cases}$

- $z_{frn} = \begin{cases} 1 & \text{if session } s \text{ is served by a n-way soft handoff with the rab } r \\ & \text{for } f \in F_s^{RL} \\ 0 & \text{otherwise.} \end{cases}$

- $w_{irfn} = \begin{cases} 1 & \text{if session } s \text{ is served by base station } i, \text{ in a n-way soft handoff} \\ & \text{with the RAB } r \text{ for } f \in F_s^{RL} \\ 0 & \text{otherwise.} \end{cases}$

- C_ℓ : the continuous variable for the capacity of the wired link ℓ in the core network.

- $C_{\ell_{\mathrm{BS}i}}$: the continuous variable for the capacity of the wired link ℓ which connects the base station i to its corresponding BSC.

5.3 Objective function

The objective of dimensioning is to minimize the resources. In our case resources contain capacity C_ℓ of the wired links ℓ in the core network, capacity $C_{\ell_{\mathrm{BS}i}}$ of the wired links and the number of base stations $\mathcal{B}S_i$ in the radio network. We assume the capacity costs to be proportional to the capacities and lengths (denoted d_ℓ) of the wired links. Therefore, the objective can be written as

$$\min\left(w_c \sum_{\ell=1}^{m} d_\ell\, C_\ell + w_c \sum_{i=1}^{n_{\mathrm{BS}}} d_{\ell_{\mathrm{BS}i}} C_{\ell_{\mathrm{BS}i}} + w_{\mathcal{B}S} \sum_{i=1}^{n_{\mathrm{BS}}} y_i^{\mathrm{BS}} \right),$$

where w_c is the cost of a link per unit of bandwidth and per unit of length, and $w_{\mathcal{B}S}$ is the cost of one base station.

5.4 Constraints

5.4.1 Call Admission Control

For each session s there exists a set of flows $F_s = \{f_s^{\mathrm{UL\_WL}},\, f_s^{\mathrm{UL\_RL}},\, f_s^{\mathrm{DL\_WL}},\, f_s^{\mathrm{DL\_RL}}\}$ associated with that session which shows the flow in the core network and radio network. There is also a set of possible paths $V_s {=} \{V_{f_s^{\mathrm{UL\_WL}}},\, V_{f_s^{\mathrm{DL\_WL}}}\}$ which can serve the session from the source to the destination. Each path is a set of links from a BSC to the destination which is an external node.

➢ **Real-time sessions (CBR):**

- On wired links:

$$x^v_{f_s^{\mathrm{UL\_WL}}} = \beta_{v f_s^{\mathrm{UL\_WL}}}\, \rho_{f_s^{\mathrm{UL\_WL}}} \qquad v \in V_{f_s^{\mathrm{UL\_WL}}}, \tag{5.2}$$

$$x^v_{f_s^{\mathrm{DL\_WL}}} = \beta_{v f_s^{\mathrm{DL\_WL}}}\, \rho_{f_s^{\mathrm{DL\_WL}}} \qquad v \in V_{f_s^{\mathrm{DL\_WL}}}. \tag{5.3}$$

- On radio links:

$$x_{f_s^{\text{UL\_RL}}} = y_s \, t_{f_s^{\text{UL\_RL}}}, \tag{5.4}$$

$$x_{f_s^{\text{DL\_RL}}} = y_s \, t_{f_s^{\text{DL\_RL}}}. \tag{5.5}$$

➢ **Non-real-time sessions (VBR):**

- On wired links:

$$\beta_{vf_s^{\text{UL\_WL}}} \, \underline{\rho}_{f_s^{\text{UL\_WL}}} \leq x^v_{f_s^{\text{UL\_WL}}} \leq \beta_{vf_s^{\text{UL\_WL}}} \, \overline{\rho}_{f_s^{\text{UL\_WL}}} \qquad v \in V_{f_s^{\text{UL\_WL}}}, \tag{5.6}$$

$$\beta_{vf_s^{\text{DL\_WL}}} \, \underline{\rho}_{f_s^{\text{DL\_WL}}} \leq x^v_{f_s^{\text{DL\_WL}}} \leq \beta_{vf_s^{\text{DL\_WL}}} \, \overline{\rho}_{f_s^{\text{DL\_WL}}} \qquad v \in V_{f_s^{\text{DL\_WL}}}. \tag{5.7}$$

- On radio links:

$$y_s \, \underline{t}_{f_s^{\text{UL\_RL}}} \leq x_{f_s^{\text{UL\_RL}}} \leq y_s \, \overline{t}_{f_s^{\text{UL\_RL}}}, \tag{5.8}$$

$$y_s \, \underline{t}_{f_s^{\text{DL\_RL}}} \leq x_{f_s^{\text{DL\_RL}}} \leq y_s \, \overline{t}_{f_s^{\text{DL\_RL}}}. \tag{5.9}$$

The two following constraints imply that for each accepted session just one path can be selected.

$$\sum_{v \in V_{f_s^{\text{UL\_WL}}}} \beta_{vf_s^{\text{UL\_WL}}} = y_s, \tag{5.10}$$

$$\sum_{v \in V_{f_s^{\text{DL\_WL}}}} \beta_{vf_s^{\text{DL\_WL}}} = y_s. \tag{5.11}$$

It is obvious that the constraints (5.2), (5.3), (5.6) and (5.7) force a x_f^v variable to 0, when the path v is not selected. Therefore the y_s and other variables corresponding to the flows will become equal to 0 as well and the session cannot be accepted. It means that a session can be accepted if and only if the uplink and downlink flows both in core and radio networks are accepted. In this case the assigned bandwidths to these flows respect the lower and upper bounds.

On the other hand, for a given direction (uplink or downlink) if none of the paths could be selected among the set of possible paths to a determined destination, then $\sum_{v\in V_{f_s^{\text{UL\_WL}}}} \beta_{vf_s^{\text{UL\_WL}}} = 0$. Referring to (5.10) and (5.11) consequently the y_s will be equal to zero. That causes the variables $x^v_{f_s^{\text{UL\_WL}}}$ or $x^v_{f_s^{\text{DL\_WL}}}$ to be equal to zero too.

In addition as mentioned above, according to the constraints (5.2), (5.3), (5.6) and (5.7) for each session toward a destination only one path can be chosen among the set of candidate paths, therefore the variables $x^v_{f_s^{\text{UL\_WL}}}$ and $x^v_{f_s^{\text{DL\_WL}}}$ of the other paths (which are not selected) will be automatically equal to zero.

5.4.2 Link between wired part and radio part

In this model the link between wired and radio parts happens on the base station level. Recalling the variable x^i_f in the case of CBR, we have:

$$x^i_{f_s^{\text{DL\_WL}}} = \rho_{f_s^{\text{DL\_WL}}} \sum_{r\in T_s^{DL}} \alpha_{irf_s^{\text{DL\_RL}}} \qquad s \in S^{CBR}, \quad i \in L_s^{BS}, \tag{5.12}$$

$$x^i_{f_s^{\text{UL\_WL}}} = \rho_{f_s^{\text{UL\_WL}}} \sum_{r\in T_s^{UL}} \alpha_{irf_s^{\text{UL\_RL}}} \qquad s \in S^{CBR}, \quad i \in L_s^{BS}. \tag{5.13}$$

In the case of VBR we have:

$$\sum_{r\in T_s^{UL}} r_t \alpha_{irf_s^{\text{UL\_RL}}} \leq x^i_{f_s^{\text{UL\_WL}}} \leq K_1 \sum_{r\in T_s^{UL}} r_t \alpha_{irf_s^{\text{UL\_RL}}} \quad s \in S^{VBR}, \quad i \in L_s^{BS}. \tag{5.14}$$

If for a session s the base station i is not selected, $\alpha_{irf_s^{\text{UL\_RL}}}$ would be equal to zero for all r. Consequently, this forces the amount of $x^i_{f_s^{\text{UL\_WL}}}$ to zero too.

On the other hand, if the base station i is selected this constraint makes sure that the data rate on the wired link section stays greater or equal to the chosen data rate of RAB (r_t) in radio link section. This prevents the flow accumulation on the selected base stations.

$$0 \leq x^i_{f_s^{\text{DL\_WL}}} \leq \sum_{r \in T_s^{\text{DL}}} r_t \alpha_{ir f_s^{\text{DL\_RL}}} \quad s \in S^{VBR}, \quad i \in L_s^{BS}. \tag{5.15}$$

Like the uplink direction, the $x^i_{f_s^{\text{DL\_WL}}}$ is equal to zero if the base station i is not selected, but unlike the uplink case we want to force the data rate on the wired link part to be less than the data rate of the chosen RAB in the radio link part in order to prevent the accumulation of flow on the selected base stations.
Of course there should be a minimum amount for this data rate and this limit can be satisfied by delay constraint which imposes a lower bound for the $x^i_{f_s^{\text{DL\_WL}}}$.

5.4.3 Quality of Service: Delay

Referring to the discussion given in previous chapter in downlink direction for a given session we have:

$$D_s^{\text{DL\_}total} = D_s^{\text{DL\_WL}} + D_s^{\text{UL\_RL}}. \tag{5.16}$$

The delay constraint on the wired links helps to take the minimum delay for the flow of session, therefore it can be applied on the VBR sessions. The constraint (5.15) does not determine a lower limit for $x^i_{f_s^{\text{DL\_WL}}}$ while the bandwidth of the selected RAB r determines its upper limit. For the worst case we consider that $D^{max-transe,rt}_{f_s^{\text{DL\_WL}}} = D^{transe,rt}_{f_s^{\text{DL\_WL}}}$ and $D^{queuing}_{f_s^{\text{DL\_WL}}} = D_s^*$.

Referring to (4.30) and (5.16) for the sum of the delay on the radio and wired parts we have:

$$D_s^{\text{DL\_}total} = D^{max-transe,rt}_{f_s^{\text{DL\_WL}}} + D^{propag}_{f_s^{\text{DL\_WL}}} + D^{trait}_{f_s^{\text{DL\_WL}}} + \frac{\sigma_s + m_{fDL} \times \text{size of packet}}{x_{f_s^{\text{DL\_WL}}}} + D^r_{f_s^{\text{DL\_RL}}}. \tag{5.17}$$

If we consider the $D_s^{\text{DL\_MAX}}$ as the maximum delay for an IP packet on the downlink direction we will have:

$$D_s^{\text{DL\_}total} \leq D_s^{\text{DL\_MAX}}. \tag{5.18}$$

Replacing the size of packet by L^{max} the previous inequality, (5.17), changes to:

$$x_{f_s^{\text{DL\_WL}}} \geq \frac{m_{f_{\text{DL}}}^{max} L_{fDL}^{max} + \sigma_s}{D_s^{\text{DL\_MAX}} - D_{f_s^{\text{DL\_WL}}}^{max-transe,rt} - D_{f_s^{\text{DL\_WL}}}^{max-propag} - D_{f_s^{\text{DL\_WL}}}^{max-trait} - D_{f_s^{\text{DL\_RL}}}^{r}}. \tag{5.19}$$

Considering $x_{f_s^{\text{DL\_WL}}}^i$ instead of $x_{f_s^{\text{DL\_WL}}}$ by using $\alpha_{ir f_s^{\text{DL\_RL}}}$ will make the constraint more dependent to the choice of the RAB r. According to (4.22) and (4.34) both $D_{f_s^{\text{DL\_RL}}}^{\text{max-transe}}$ and $D_{f_s^{\text{DL\_RL}}}^{r}$ depend on the selection of the RAB r, therefore we conclude:

$$x_{f_s^{\text{DL\_WL}}}^i \geq \sum_{r \in T_s^{DL}} \frac{m_{f_{\text{DL}}}^{max} L_{f_{\text{DL}}}^{max} + \sigma_s}{D_s^{\text{DL\_MAX}} - D_{f_s^{\text{DL\_WL}}}^{max-transe,rt} - D_{f_s^{\text{DL\_WL}}}^{max-propag} - D_{f_s^{\text{DL\_WL}}}^{max-trait} - D_{f_s^{\text{DL\_RL}}}^{r}} \alpha_{ir f_s^{\text{DL\_RL}}}$$

$$s \in S^{\text{VBR}}, \quad i \in L_s^{\text{BS}}. \tag{5.20}$$

Recalling the $\mathcal{R}(r_t)$ as the minimum data rate for the flow $f_s^{\text{DL-WL}}$ if the session is served by the data rate r_t on the radio link the formula (5.20) changes to (5.21) in order to respect the lower limits:

$$x_{f_s^{\text{DL\_WL}}}^i \geq$$

$$\sum_{r \in T_s^{DL}} \max(\mathcal{R}(r_t), \frac{m_{f_{\text{DL}}}^{max} L_{f_{\text{DL}}}^{max} + \sigma_s}{D_s^{\text{DL\_MAX}} - D_{f_s^{\text{DL\_WL}}}^{max-transe,rt} - D_{f_s^{\text{DL\_WL}}}^{max-propag} - D_{f_s^{\text{DL\_WL}}}^{max-trait} - D_{f_s^{\text{DL\_RL}}}^{r}}) \alpha_{ir f_s^{\text{DL\_RL}}}$$

$$s \in S^{\text{VBR}}, \quad i \in L_s^{\text{BS}}. \tag{5.21}$$

In this model, the $m_{f_{\text{DL}}}$ is not fixed, as we may have different possible paths with different number of links. Therefore we assume that $m_{f_{\text{DL}}}^{max}$ is always equal to the links number of the longest path in the set of possible paths. That is the reason why we have considered the maximum amount for all the different types of delay.

5.4.4 Capacity of wired links

Referring to Chapter 3, there are two types of wired links in the proposed network model. First the wired links in radio network which are the links that connect each base station to its BSC and second the links in the core network which define the selected path for each session. Both the serving base station (or base stations in the soft handoff cases) and selected path for a session are not fixed from the beginning. The optimization procedure chooses a base station or more and a path for each accepted session among the set of potential base stations and possible paths. Note that the potential base stations can be connected to different BSCs. We discus the constraints on each type of wired links one after the other.

Capacity of wired links in radio network

For each link between a base station and its corresponding BSC there exists a wired link and we call the capacity of this link as $C_{\ell_{BS_i}}$.

$$\sum_{s\in S}\sum_{f\in F_s^{\mathrm{WL}}} b_{ihf}x_f^i \leq C_{\ell_{BS_i}} \qquad h\in H, \quad i\in L^{\mathrm{BS}}. \tag{5.22}$$

For a session the x_f^i is equal to x_f if the base station i serves that session. Therefore the session contributes to the dimensioning of the wired link between the base station i and its BSC. While if $x_f^i = 0$ it means that base station i is not selected to serve that session, so it does not play any role in the dimensioning.

Capacity of wired links in core network

The objective function aims to minimize the capacity, C_ℓ. Therefore, for the set of possible paths of a session which contain link ℓ, the sum of flow on each period h that pass link ℓ should be less than the capacity C_ℓ. Since the total bandwidth on a link cannot exceed the capacity of this link, we have considered the following constraint.

$$\sum_{s \in S} \sum_{v \in V_{f_s^{\text{UL\_WL}}}} a_{hs}\, m_{vl}\, x^{v}_{f_s^{\text{UL\_WL}}} + \sum_{s \in S} \sum_{v \in V_{f_s^{\text{DL\_WL}}}} a_{hs}\, m_{vl}\, x^{v}_{f_s^{\text{DL\_WL}}} \leq C_{\ell}$$

$$h \in H, \qquad \ell = 1, 2, ..., m. \quad (5.23)$$

As the links are used in uplink and downlink directions therefore the sum of flows in both directions have been considered.

5.4.5 Capacity of radio links

In the radio link, during each period of time, the total used radio capacity by the sessions in each direction should be less than the maximum capacity of the base station which gives service to those sessions. This definition is studied in Section 4.1.4 for the uplink and downlink directions individually. We also discussed the soft handoff gain which reduces the necessary radio capacity in the soft handoff cases.

Radio Capacity on the uplink direction

$$\frac{1+\lambda}{W} \sum_{s \in S} \sum_{r \in T_{f_s^{\text{UL\_RL}}}} \Bigg(\frac{\gamma_{sr}^{\text{UL}}\ \nu_s\ b_{ihf_s^{\text{UL\_RL}}}\, \alpha_{irf_s^{\text{UL\_RL}}}}{\frac{1}{r_t} + \frac{\gamma_{sr}^{\text{UL}}\, \nu_s}{W}} - \sum_{n=2}^{U} \frac{\gamma_{sr}^{\text{UL}}\ \nu_s\ b_{ihf_s^{\text{UL\_RL}}}\, (1 - SHG_n^{\text{UL}}) w_{irf_s^{\text{UL\_RL}}n}}{(1 + \frac{\gamma_{sr}^{\text{UL}}\, \nu_s\, r_t}{W})\, (\frac{1}{r_t} + \frac{\gamma_{sr}^{\text{UL}}\, \nu_s\, SHG_n^{\text{UL}}}{W})} \Bigg) \leq L_{\text{UL}}$$

$$i \in L^{\text{BS}}, \quad h \in H. \quad (5.24)$$

Considering that session s is accepted in soft handoff with two base stations and one of them is base station i then both $\alpha_{irf_s^{\text{UL\_RL}}}$ and $w_{irf_s^{\text{UL\_RL}}2}$ would be equal to one. It means that the session s contributes in the calculation of the total capacity for the base station i. Therefore the capacity occupied by the session s (C_s) would be equal to:

$$C_s = \frac{1+\lambda}{W} \frac{\gamma_{sr}^{\text{UL}}\ \nu_s\ b_{ihf_s^{\text{UL\_RL}}} \times 1}{\frac{1}{r_t} + \frac{\gamma_{sr}^{\text{UL}}\, \nu_s}{W}}$$

$$-\frac{\gamma_{sr}^{\text{UL}}\ \nu_s\ b_{ihf_s^{\text{UL\_RL}}}\,(1-SHG_2^{\text{UL}})\times 1}{(1+\frac{\gamma_{sr}^{\text{UL}}\,\nu_s\,r_t}{W})\,(\frac{1}{r_t}+\frac{\gamma_{sr}^{\text{UL}}\ \nu_s\,SHG_2^{\text{UL}}}{W})}.$$

If the session is accepted in soft handoff with three base stations then we would have $w_{irf_s^{\text{UL\_RL}}3}=1$, therefore the occupied capacity by the session s would be equal to:

$$C_s=\frac{1+\lambda}{W}\frac{\gamma_{sr}^{\text{UL}}\ \nu_s\ b_{ihf_s^{\text{UL\_RL}}}\times 1}{\frac{1}{r_t}+\frac{\gamma_{sr}^{\text{UL}}\,\nu_s}{W}}$$

$$-\frac{\gamma_{sr}^{\text{UL}}\ \nu_s\ b_{ihf_s^{\text{UL\_RL}}}\,(1-SHG_3^{\text{UL}})\times 1}{(1+\frac{\gamma_{sr}^{\text{UL}}\,\nu_s\,r_t}{W})\,(\frac{1}{r_t}+\frac{\gamma_{sr}^{\text{UL}}\ \nu_s\,SHG_3^{\text{UL}}}{W})}.$$

On the other hand, if session s is accepted but not in soft handoff then $w_{irf_s^{\text{UL\_RL}}n}=0$ while $\alpha_{irf_s^{\text{UL\_RL}}}=1$ and the radio capacity occupied by this session would be equal to:

$$C_s=\frac{1+\lambda}{W}\frac{\gamma_{sr}^{\text{UL}}\ \nu_s\ b_{ihf_s^{\text{UL\_RL}}}\times 1}{\frac{1}{r_t}+\frac{\gamma_{sr}^{\text{UL}}\,\nu_s}{W}}.$$

Obviously, if session s is not taken by the base station i then both $\alpha_{irf_s^{\text{UL\_RL}}}$ and $w_{irf_s^{\text{UL\_RL}}n}$ would be equal to zero, so this session does not play any role in the total capacity for the base station i.

Radio Capacity on the downlink direction

$$\sum_{s\in S}\sum_{r\in T_{f_s^{\text{DL\_RL}}}}\Bigg(\frac{\gamma_{sr}^{\text{DL}}\ \nu_s\left(N_0 a_{m_s,i}+\frac{P_{\text{BS}}}{W}(\sum_{j\in Ring(BS_i)}\frac{a_{m_s,i}}{a_{m_s,j}}+w)\right)b_{ihf^{\text{DL\_RL}}-s}\ \alpha_{irf_s^{\text{DL\_RL}}}}{\frac{1}{r_t}+\frac{\gamma_{sr}^{\text{DL}}\nu_s w}{W}}$$

$$-\sum_{n=2}^{U}\frac{\gamma_{sr}^{\text{DL}}\ \nu_s\left(N_0 a_{m_s,i}+\frac{P_{\text{BS}}}{W}(\sum_{j\in Ring(BS_i)}\frac{a_{m_s,i}}{a_{m_s,j}}+w)\right)(1-SHG_n^{\text{DL}})\,b_{ihf_s^{\text{DL\_RL}}}\ w_{irf_s^{\text{DL\_RL}}n}}{(1+\frac{\gamma_{sr}^{\text{DL}}\,\nu_s\,w\,r_t}{W})\,(\frac{1}{r_t}+\frac{\gamma_{sr}^{\text{DL}}\,\nu_s\,w\,SHG_n^{\text{DL}}}{W})}\Bigg)$$

$$\leq P_{\text{BS}}-P_{cont}$$

$$i\in L^{\text{BS}},\quad h\in H.\quad (5.25)$$

The same procedures done for the uplink can be applied for the downlink direction to calculate the occupied capacity by a session in soft handoff or non soft handoff.

5.4.6 Grade of Service

$$\frac{\sum\limits_{s \in S^{a,p}} y_s}{|S^{a,p}|} \geq 1 - \mathrm{Br}_{a,p} \qquad a \in A \quad p \in P. \tag{5.26}$$

For each group of priority (Gold, Silver) and each kind of application (voice, videophone, video streaming, web browsing and mail) there is a threshold on the number of accepted sessions during the planning time, which depends on the rate of blocking for flows.

5.4.7 Handshake between a BSC and a base station

Recall that

$$x^i_{f^{\text{WL}}} = \begin{cases} x_{f^{\text{WL}}} & \text{if the base station } i \text{ has been selected for session } s \\ 0 & \text{otherwise.} \end{cases}$$

The handshake deals with the uplink and downlink flows at the BSC. In the radio link for each session there is a list of potential base stations which can carry the session. We call this list L^{BS}_s. These base stations may be connected to a unique or different BSCs. N_{BS_k} is the notation that we use to call the set of potential base stations connected to BSC_k with fixed and known geographical positions.

The data rate of flow(s) in link(s) of a selected path in the core network would be equal to the data rate of flow in all links to base station(s) in the radio network. It means that if we call the flow corresponding to a path in the uplink direction as $x^v_{f_s^{\text{UL\_WL}}}$ it would be equal to $x^i_{f_s^{\text{UL\_WL}}}$ if base station i and path v are selected to serve session s. The complexity shows itself in the soft handoff situations where we would have $x^v_{f_s^{\text{UL\_WL}}}$ equal to the sum of two or three coming flows of two or three different base stations.

The following constraints show the remedy of this complexity in the uplink and downlink directions.

$$(1+\sum_{n=2}^{U}\sum_{r\in T_s^{\mathrm{UL}}}(n-1)z_{f_s^{\mathrm{UL\_RL}}rn})\sum_{v\in V^k_{f_s^{\mathrm{UL\_WL}}}} x^v_{f_s^{\mathrm{UL\_WL}}} = \sum_{i\in L_s^{BS}\cap N_{BS_k}} x^i_{f_s^{\mathrm{UL\_WL}}} \qquad s\in S, \quad k\in K_s, \tag{5.27}$$

$$(1+\sum_{n=2}^{U}\sum_{r\in T_s^{\mathrm{DL}}}(n-1)z_{f_s^{\mathrm{DL\_RL}}rn})\sum_{v\in V^k_{f_s^{\mathrm{DL\_WL}}}} x^v_{f_s^{\mathrm{DL\_WL}}} = \sum_{i\in L_s^{BS}\cap N_{BS_k}} x^i_{f_s^{\mathrm{DL\_WL}}} \qquad s\in S, \quad k\in K_s. \tag{5.28}$$

Even in the soft handoff case for each session just one and only one path between BSC and destination will be accepted. The former proposed constraints are not linear, so we replace them with the following linear constraints:

$$x^i_{f_s^{\mathrm{UL\_WL}}} \le \sum_{v\in V^k_{f_s^{\mathrm{UL\_WL}}}} x^v_{f_s^{\mathrm{UL\_WL}}} \le \sum_{j\in L_s^{BS}\cap N_{BS_k}} x^j_{f_s^{\mathrm{UL\_WL}}} \qquad s\in S, \quad k\in K_s, \quad i\in L_s^{BS}\cap N_{BS_k}, \tag{5.29}$$

$$x^i_{f_s^{\mathrm{DL\_WL}}} \le \sum_{v\in V^k_{f_s^{\mathrm{DL\_WL}}}} x^v_{f_s^{\mathrm{DL\_WL}}} \le \sum_{j\in L_s^{BS}\cap N_{BS_k}} x^j_{f_s^{\mathrm{DL\_WL}}} \qquad s\in S, \quad k\in K_s, \quad i\in L_s^{BS}\cap N_{BS_k}. \tag{5.30}$$

The constraints (5.29) and (5.30) are, however, not equivalent to (5.27) and (5.28), so some optimal solutions may not satisfy $x^i_f = x^v_f$ if base station i and path v are selected, but it is always possible to construct another optimal solution with same value that satisfies this equality.

5.4.8 Soft handoff

In this thesis we have studied only the micro diversity where the potential base stations which can serve a session in soft handoff are all connected to the same BSC. According to the definition of z_{frn} if a session is accepted as a soft handoff then we are going to have $z_{frn} = 1$ if it is a n-way soft handoff, and obviously $z_{frn} = 0$ when the session is not a soft handoff or when it is refused.

Therefore we will have:

$$\sum_{n=2}^{U} \sum_{r \in T_s^{\text{UL}}} z_{f_s^{\text{UL\_RL}} rn} \leq y_s \qquad s \in S, \tag{5.31}$$

$$\sum_{n=2}^{U} \sum_{r \in T_s^{\text{DL}}} z_{f_s^{\text{DL\_RL}} rn} \leq y_s \qquad s \in S. \tag{5.32}$$

Recalling the definition of α_{irf} if the session is non-soft handoff then

$$\sum_{i \in L_s^{BS}} \sum_{r \in T_s} \alpha_{irf} = 1. \tag{5.33}$$

If the session is soft handoff with n base stations then

$$\sum_{i \in L_s^{BS}} \sum_{r \in T_s} \alpha_{irf} = n. \tag{5.34}$$

Since we should have exactly $n\alpha_{irf}$ variable equal to 1 in the n-way soft handoff, we propose the following constraints:

$$y_s = \sum_{i \in L_s^{\text{BS}}} \sum_{r \in T_s^{\text{UL}}} \alpha_{irf_s^{\text{UL\_RL}}} - \sum_{n=2}^{U} \sum_{r \in T_s^{\text{UL}}} (n-1)\, z_{f_s^{\text{UL\_RL}} rn} \qquad s \in S, \tag{5.35}$$

$$y_s = \sum_{i \in L_s^{\text{BS}}} \sum_{r \in T_s^{\text{DL}}} \alpha_{irf_s^{\text{DL\_RL}}} - \sum_{n=2}^{U} \sum_{r \in T_s^{\text{DL}}} (n-1)\, z_{f_s^{\text{DL\_RL}} rn} \qquad s \in S. \tag{5.36}$$

Recalling the definition of w_{irfn} if the session is a n-way soft handoff therefore

$$\sum_{i \in L_s^{BS-SH}} w_{irfn} = n, \tag{5.37}$$

but in all cases the z_{frn} would be always equal to one. In order to force this specification we should satisfy the following constraints:

$$z_{f_s^{\mathrm{UL\_RL}}rn} = \frac{1}{n} \sum_{i \in L_s^{BS-SH}} w_{irf_s^{\mathrm{UL\_RL}}n} \qquad r \in T_s, \quad s \in S, \quad n = 2, .., U, \tag{5.38}$$

$$z_{f_s^{\mathrm{DL\_RL}}rn} = \frac{1}{n} \sum_{i \in L_s^{BS-SH}} w_{irf_s^{\mathrm{DL\_RL}}n} \qquad r \in T_s, \quad s \in S, \quad n = 2, .., U. \tag{5.39}$$

On the other hand, we should always respect the following relation between the variables α_{irf} and w_{irfn}.

$$\alpha_{irf_s^{\mathrm{UL\_RL}}} \geq \sum_{n=2}^{U} w_{irf_s^{\mathrm{UL\_RL}}n} \qquad r \in T_s, \quad s \in S, \quad i \in L_s^{BS-SH}, \tag{5.40}$$

$$\alpha_{irf_s^{\mathrm{DL\_RL}}} \geq \sum_{n=2}^{U} w_{irf_s^{\mathrm{DL\_RL}}n} \qquad r \in T_s, \quad s \in S, \quad i \in L_s^{BS-SH}. \tag{5.41}$$

In this level we need a constraint to avoid the appearance of soft handoff between the potential base stations connected to the different BSC. In order to satisfy this idea the following constraints should be considered for each session to make sure that potential base stations which can serve the session in soft handoff are not connected to the same BSC.

$$w_{irf_s^{\mathrm{UL\_RL}}n} + w_{jrf_s^{\mathrm{UL\_RL}}n} \leq 1 \qquad r \in T_s^{\mathrm{UL}}, \quad n = 2, .., U, \quad i, j \in L_s^{BS-SH}, \tag{5.42}$$

$$w_{irf_s^{\mathrm{DL\_RL}}n} + w_{jrf_s^{\mathrm{DL\_RL}}n} \leq 1 \qquad r \in T_s^{\mathrm{DL}}, \quad n = 2, .., U, \quad i, j \in L_s^{BS-SH}. \tag{5.43}$$

If we consider r_1 and n_1 as a subset of r and n then for sure we will have

$$w_{ir_1 f_s^{\text{UL\_RL}} n_1} + w_{jr_1 f_s^{\text{UL\_RL}} n_1} \leq \sum_{n=2}^{U} \sum_{r \in T_s^{\text{UL}}} (w_{irf_s^{\text{UL\_RL}} n} + w_{jrf_s^{\text{UL\_RL}} n}) \leq 1. \tag{5.44}$$

It is also obvious that if a session s is accepted in a n-way soft handoff mode there exists at most one couple (r,n) such that $w_{irf_s^{\text{UL\_RL}} n}$ can be non zero.
Therefore if we have

$$w_{irf_s^{\text{UL\_RL}} n} + w_{jrf_s^{\text{UL\_RL}} n} \leq 1, \tag{5.45}$$

then,

$$w_{ir_1 f_s^{\text{UL\_RL}} n_1} + w_{jr_1 f_s^{\text{UL\_RL}} n_1} = 0. \tag{5.46}$$

Consequently the inequality (5.44) can be transformed to the following constraint.

$$\sum_{n=2}^{U} \sum_{r \in T_s^{\text{UL}}} (w_{irf_s^{\text{UL\_RL}} n} + w_{jrf_s^{\text{UL\_RL}} n}) \leq 1$$

$$s \in S, \quad r \in T_s^{\text{DL}}, \quad n = 2, .., U, \quad i, j \in L_s^{BS-SH}, \quad BSC_i \neq BSC_j. \tag{5.47}$$

With the same reasoning for the downlink direction we have:

$$\sum_{n=2}^{U} \sum_{r \in T_s^{\text{DL}}} (w_{irf_s^{\text{DL\_RL}} n} + w_{jrf_s^{\text{DL\_RL}} n}) \leq 1$$

$$s \in S, \quad r \in T_s^{\text{DL}}, \quad n = 2, .., U, \quad i, j \in L_s^{BS-SH}, \quad BSC_i \neq BSC_j. \tag{5.48}$$

5.4.9 Quality of Service: Selection of a RAB on the radio links

A set of discrete values of RABs is considered for each session in uplink and downlink direction. Hence, in any uplink and downlink flow f on the radio network we must satisfy for each session $s \in S$:

$$x_{f_s^{\text{UL\_RL}}} = \sum_{r \in T_s^{UL}} r_t(\sum_{i \in L_s^{BS}} \alpha_{irf_s^{\text{UL\_RL}}} - \sum_{n=2}^{U} (n-1)\, z_{frn}) \quad s \in S, \tag{5.49}$$

$$x_{f_s^{\mathrm{DL\_RL}}} = \sum_{r \in T_s^{DL}} r_t\Big(\sum_{i \in L_s^{BS}} \alpha_{irf_s^{\mathrm{DL\_RL}}} - \sum_{n=2}^{U} (n-1)\, z_{frn}\Big) \quad s \in S. \tag{5.50}$$

Obviously, $x_{f_s^{\mathrm{UL\_RL}}}$ and $x_{f_s^{\mathrm{DL\_RL}}}$ obtain the data rate of RAB r if this RAB is selected. $\sum_{i \in L_s^{BS}} \alpha_{irf} - \sum_{n=2}^{U} (n-1)\, z_{frn}$ defines a binary variable which is equal to 1 if RAB r is selected even in soft handoff or non soft handoff and is equal to zero if it is not selected.

Now we should have a constraint which forces the selection of the same base station both for uplink and downlink.

$$\sum_{r \in T_s^{\mathrm{UL}}} \alpha_{irf_s^{\mathrm{UL\_RL}}} = \sum_{r \in T_s^{\mathrm{DL}}} \alpha_{irf_s^{\mathrm{DL\_RL}}} \qquad s \in S, i \in L_s^{BS}. \tag{5.51}$$

As mentioned in Section 4.2 for each application we introduce two different priorities named Gold and Silver. Obviously, a Gold session has a higher priority than a Silver one. On the other hand, for each pair of (application, priority) e.g., (web browsing, Gold) in each direction e.g., (downlink) there is a set of RABs, $\{r_1, r_2, ..., r_n\}$ with n coefficients $Q_{r_1}, Q_{r_2}, ..., Q_{r_n}$. Coefficient Q_{r_1} represents the rate of (web browsing, Gold) sessions which should be accepted with RAB number 1 in downlink direction. Considering Table 5.1 where four different RABs are available we have:

ℓ	r_ℓ	Q_{r_ℓ}	Q_{r_ℓ}
		Gold	Silver
1	76.8, 5%	0.6	0.1
2	76.8, 10%	0.25	0.05
3	38.4, 5%	0.1	0.6
4	38.4, 10%	0.05	0.25

TAB. 5.1: Value of the QoS coefficients for web browsing downlink

$$\sum_{\ell=1}^{4} Q_{r_\ell} = 1.$$

If Y is the total number of accepted (web browsing, Gold) sessions and Y_ℓ represents the number of accepted sessions with the Rab ℓ then:

$$\sum_{\ell=1}^{4} Y_\ell = Y.$$

Therefore for RAB r_1

$$Q_{r_1} Y - \sigma \leq Y_1 \leq Q_{r_1} Y + 1 - \sigma,$$

where σ is a parameter in the [0,1) interval.

The idea is to observe that the value of QoS $\frac{Y_1}{Y}$ lies in an interval of length 1 containing the value of Q_{r_1}. The same reasoning applies for the second RAB on the remaining sessions. Where $\acute{Q}_{r_2} = \frac{Q_{r_2}}{1-Q_{r_1}}$ then $\sum_{r=1}^{n} \acute{Q}_{r_2} = 1$. We also assume $\sigma = \frac{1}{2}$ and

$$\alpha_{rf} = \sum_{i \in L_s^{\text{BS}}} \alpha_{irf} - \sum_{n=2}^{U} (n-1)\, z_{frn} \qquad r \in T_f, \quad f \in F_s^{\text{RL}}, \quad s \in S,$$

therefore we will have:

$$\frac{Q^{\text{UL}}_{r_\ell,a,p}}{\sum_{j=\ell}^{|T^{\text{UL}}_{a,p}|} Q^{\text{UL}}_{r_j,a,p}} \sum_{s \in S^{a,p}} \sum_{j=\ell+1}^{|T^{\text{UL}}_{a,p}|} \alpha_{r_j f_s^{\text{UL\_RL}}} - \frac{1}{2} \leq \left(1 - \frac{Q^{\text{UL}}_{r_\ell,a,p}}{\sum_{j=\ell}^{|T^{\text{UL}}_{a,p}|} Q^{\text{UL}}_{r_j,a,p}} \right) \sum_{s \in S^{a,p}} \alpha_{r_\ell f_s^{\text{UL\_RL}}}$$

$$\leq \frac{Q^{\text{UL}}_{r_\ell,a,p}}{\sum_{j=\ell}^{|T^{\text{UL}}_{a,p}|} Q^{\text{UL}}_{r_j,a,p}} \sum_{s \in S^{a,p}} \sum_{j=\ell+1}^{|T^{\text{UL}}_{a,p}|} \alpha_{r_j f_s^{\text{UL\_RL}}} + \frac{1}{2}$$

$$(r_1, r_2, \ldots) \in T^{\text{UL}}_{a,p}, \quad \ell = 1, \ldots, |T^{\text{UL}}_{a,p}| - 1, \quad a \in A, \quad p \in P. \quad (5.52)$$

$$\frac{Q^{\text{DL}}_{r_\ell,a,p}}{\sum\limits_{j=\ell}^{|T^{\text{DL}}_{a,p}|} Q^{\text{DL}}_{r_j,a,p}} \sum_{s\in S^{a,p}} \sum_{j=\ell+1}^{|T^{\text{DL}}_{a,p}|} \alpha_{r_j f_s^{\text{DL\_RL}}} - \frac{1}{2} \leq \left(1 - \frac{Q^{\text{DL}}_{r_\ell,a,p}}{\sum\limits_{j=\ell}^{|T^{\text{DL}}_{a,p}|} Q^{\text{DL}}_{r_j,a,p}} \right) \sum_{s\in S^{a,p}} \alpha_{r_\ell f_s^{\text{DL\_RL}}}$$

$$\leq \frac{Q^{\text{DL}}_{r_\ell,a,p}}{\sum\limits_{j=\ell}^{|T^{\text{DL}}_{a,p}|} Q^{\text{DL}}_{r_j,a,p}} \sum_{s\in S^{a,p}} \sum_{j=\ell+1}^{|T^{\text{DL}}_{a,p}|} \alpha_{r_j f_s^{\text{DL\_RL}}} + \frac{1}{2}$$

$$(r_1, r_2, \ldots) \in T^{\text{DL}}_{a,p}, \quad \ell = 1, \ldots, |T^{\text{DL}}_{a,p}| - 1, \quad a \in A, \quad p \in P. \quad (5.53)$$

Note that the RABs are sorted in the way that the best one is the first. A RAB r_1 is said to be better than the RAB r_2 if $r_{2t} \leq r_{1t}$. If $r_{2t} = r_{1t}$ then the better RAB would be the one with less Frame Error Rate.

5.4.10 Activation of a base station

If a base station i has been selected for at least one session, we must activate that base station. The following constraints make sure that base station i is activated if there is even one $\alpha_{irf_s^{\text{DL-RL}}}$ which is equal to one.

$$y_i^{\text{BS}} \leq \sum_{s\in S} \sum_{r\in T_{f_s^{\text{DL\_RL}}}} \alpha_{irf_s^{\text{DL-RL}}} \qquad i \in L^{\text{BS}}, \qquad (5.54)$$

$$\sum_{r\in T_{f_s^{\text{DL}}}} \alpha_{irf_s^{\text{DL-RL}}} \leq y_i^{\text{BS}} \qquad i \in L_s^{\text{BS}}, \quad s \in \mathrm{S}. \qquad (5.55)$$

As mentioned before there are five different types of applications. All of them may use the downlink and uplink direction for the same session. Among them just the video streaming never use the uplink direction. The mail application, however, may use the uplink or downlink direction one in a time. So we limit the use of the uplink sense of above constraints just for mail. Recall that $\sum\limits_{r\in T_{f_s^{\text{DL}}}} \alpha_{irf_s^{\text{DL-RL}}} = \sum\limits_{r\in T_{f_s^{\text{UL}}}} \alpha_{irf_s^{\text{UL-RL}}}$, in this case, the $\sum\limits_{r\in T_{f_s^{\text{DL}}}} \alpha_{irf_s^{\text{DL-RL}}}$ will be replaced by $\sum\limits_{r\in T_{f_s^{\text{UL}}}} \alpha_{irf_s^{\text{UL-RL}}}$.

5.5 Bounds and domains of the variables: Summing up

Flow variables on the wired link:

$x_f \in \mathbb{R}^+ \qquad f \in F \qquad s \in S.$

Duplicated flow variables (caused by soft handoff, or selection of an active BS):

$x_f^i \in \mathbb{R}^+ \qquad f \in F_s^{\text{WL}}, \qquad i \in L_s^{\text{BS}}, \qquad s \in S.$

Flow variables on each link of the wired links:

$x_f^v \in \mathbb{R}^+ \qquad f \in F_s^{\text{WL}}, \qquad v \in V_f, \qquad s \in S.$

Call admission control variables:

$y_s \in \{0,1\} \qquad s \in S.$

variable for choosing the RAB and base station:

$\alpha_{irf} \in \{0,1\} \qquad f \in F^{\text{RL}} \qquad i \in L_s^{\text{BS}}, \qquad s \in S \qquad r \in T_s^{\text{UL}} \text{ or } T_s^{\text{DL}}.$

variable for choosing the Path:

$\beta_{vf} \in \{0,1\} \qquad f \in F_s^{\text{WL}}, \qquad v \in V_f, \qquad s \in S.$

Wired capacity variables for the wired link in Core Network:

$C_\ell \in \mathbb{R}^+ \qquad \ell = 1, 2, \ldots, m.$

Wired capacity variables for the wired link in Radio Network:

$C_{\ell_{\text{BS}i}} \in \mathbb{R}^+ \qquad i \in L^{\text{BS}}.$

Active base station variables:

$y_i^{\text{BS}} \in \{0,1\} \qquad i \in L^{\text{BS}}.$

Soft handoff variables:

$z_{frn} \in \{0,1\} \quad f \in F^{\text{RL}}, \quad r \in T_s^{\text{UL}} \text{ or } T_s^{\text{DL}}, \quad n = 2, \ldots U, \quad s \in S.$

$w_{irfn} \in \{0,1\} \quad f \in F^{\text{RL}} \quad i \in L^{\text{BS\_SH}}, \quad r \in T_s^{\text{UL}} \text{ or } T_s^{\text{DL}}, \quad n = 2, \ldots, U, \quad s \in S.$

5.6 Dimensioning mathematical model: Summing up

We sum up below the set of constraints of the anticipative mathematical model as well as the objective function.

$$\min \left(w_c \sum_{\ell=1}^{m} d_\ell \, C_\ell + w_c \sum_{i=1}^{n_{\text{BS}}} d_{\ell_{\text{BS}i}} C_{\ell_{\text{BS}i}} + w_{BS} \sum_{i=1}^{n_{\text{BS}}} y_i^{\text{BS}} \right)$$

Call Admission Control:

$\forall\, s \in S$ with $F_s = \{f^{\mathrm{UL\_WL}}, f^{\mathrm{UL\_RL}}, f^{\mathrm{DL\_WL}}, f^{\mathrm{DL\_RL}}\}$

When the session is CBR:

$*(5.2) \qquad x^{v}_{f_s^{\mathrm{UL\_WL}}} = \beta_{vf_s^{\mathrm{UL\_WL}}}\, \rho_{f_s^{\mathrm{UL\_WL}}} \qquad v \in V_{f_s^{\mathrm{UL\_WL}}}$

$*(5.3) \qquad x^{v}_{f_s^{\mathrm{DL\_WL}}} = \beta_{vf_s^{\mathrm{DL\_WL}}}\, \rho_{f_s^{\mathrm{DL\_WL}}} \qquad v \in V_{f_s^{\mathrm{DL\_WL}}}$

$(5.4) \qquad x_{f_s^{\mathrm{UL\_RL}}} = y_s\, t_{f_s^{\mathrm{UL\_RL}}}$

$(5.5) \qquad x_{f_s^{\mathrm{DL\_RL}}} = y_s\, t_{f_s^{\mathrm{DL\_RL}}}$

When the session is VBR:

$*(5.6) \qquad \beta_{vf_s^{\mathrm{UL\_WL}}}\, \underline{\rho}_{f_s^{\mathrm{UL\_WL}}} \leq x^{v}_{f_s^{\mathrm{UL\_WL}}} \leq \beta_{vf_s^{\mathrm{UL\_WL}}}\, \overline{\rho}_{f_s^{\mathrm{UL\_WL}}} \qquad v \in V_{f_s^{\mathrm{UL\_WL}}}$

$*(5.7) \qquad \beta_{vf_s^{\mathrm{DL\_WL}}}\, \underline{\rho}_{f_s^{\mathrm{DL\_WL}}} \leq x^{v}_{f_s^{\mathrm{DL\_WL}}} \leq \beta_{vf_s^{\mathrm{DL\_WL}}}\, \overline{\rho}_{f_s^{\mathrm{DL\_WL}}} \qquad v \in V_{f_s^{\mathrm{DL\_WL}}}$

$(5.8) \qquad y_s\, \underline{t}_{f_s^{\mathrm{UL\_RL}}} \leq x_{f_s^{\mathrm{UL\_RL}}} \leq y_s\, \overline{t}_{f_s^{\mathrm{UL\_RL}}}$

$(5.9) \qquad y_s\, \underline{t}_{f_s^{\mathrm{DL\_RL}}} \leq x_{f_s^{\mathrm{DL\_RL}}} \leq y_s\, \overline{t}_{f_s^{\mathrm{DL\_RL}}}$

$*(5.10) \qquad \sum_{v \in V_{f_s^{\mathrm{UL\_WL}}}} \beta_{vf_s^{\mathrm{UL\_WL}}} = y_s$

$*(5.11) \qquad \sum_{v \in V_{f_s^{\mathrm{DL\_WL}}}} \beta_{vf_s^{\mathrm{DL\_WL}}} = y_s$

Grade of Service

$(5.26) \qquad \frac{\sum_{s \in S^{a,p}} y_s}{|S^{a,p}|} \geq 1 - \mathrm{Br}_{a,p} \qquad a \in A, \quad p \in P$

Quality of Service

$$* \, (5.52) \quad \frac{Q^{\text{UL}}_{r_\ell,a,p}}{\sum\limits_{j=\ell}^{|T^{\text{UL}}_{a,p}|} Q^{\text{UL}}_{r_j,a,p}} \sum_{s \in S^{a,p}} \sum_{j=\ell+1}^{|T^{\text{UL}}_{a,p}|} \alpha_{r_j f^{\text{UL\_RL}}_s} - \frac{1}{2} \leq \left(1 - \frac{Q^{\text{UL}}_{r_\ell,a,p}}{\sum\limits_{j=\ell}^{|T^{\text{UL}}_{a,p}|} Q^{\text{UL}}_{r_j,a,p}} \right) \sum_{s \in S^{a,p}} \alpha_{r_\ell f^{\text{UL\_RL}}_s}$$

$$\leq \frac{Q^{\text{UL}}_{r_\ell,a,p}}{\sum\limits_{j=\ell}^{|T^{\text{UL}}_{a,p}|} Q^{\text{UL}}_{r_j,a,p}} \sum_{s \in S^{a,p}} \sum_{j=\ell+1}^{|T^{\text{UL}}_{a,p}|} \alpha_{r_j f^{\text{UL\_RL}}_s} + \frac{1}{2}$$

$$(r_1, r_2, \dots) \in T^{\text{UL}}_{a,p}, \quad \ell = 1, \dots, |T^{\text{UL}}_{a,p}| - 1, \quad a \in A, \quad p \in P$$

$$* \, (5.53) \quad \frac{Q^{\text{DL}}_{r_\ell,a,p}}{\sum\limits_{j=\ell}^{|T^{\text{DL}}_{a,p}|} Q^{\text{DL}}_{r_j,a,p}} \sum_{s \in S^{a,p}} \sum_{j=\ell+1}^{|T^{\text{DL}}_{a,p}|} \alpha_{r_j f^{\text{DL\_RL}}_s} - \frac{1}{2} \leq \left(1 - \frac{Q^{\text{DL}}_{r_\ell,a,p}}{\sum\limits_{j=\ell}^{|T^{\text{DL}}_{a,p}|} Q^{\text{DL}}_{r_j,a,p}} \right) \sum_{s \in S^{a,p}} \alpha_{r_\ell f^{\text{DL\_RL}}_s}$$

$$\leq \frac{Q^{\text{DL}}_{r_\ell,a,p}}{\sum\limits_{j=\ell}^{|T^{\text{DL}}_{a,p}|} Q^{\text{DL}}_{r_j,a,p}} \sum_{s \in S^{a,p}} \sum_{j=\ell+1}^{|T^{\text{DL}}_{a,p}|} \alpha_{r_j f^{\text{DL\_RL}}_s} + \frac{1}{2}$$

$$(r_1, r_2, \dots) \in T^{\text{DL}}_{a,p}, \quad \ell = 1, \dots, |T^{\text{DL}}_{a,p}| - 1, \quad a \in A, \quad p \in P$$

$$\alpha_{rf} = \sum_{i \in L^{\text{BS}}_s} \alpha_{irf} - \sum_{n=2}^{U} (n-1) \, z_{frn}, \qquad r \in T_f, \quad f \in F^{\text{RL}}_s, \quad s \in S$$

Radio Link Capacity

$$x_{f_s^{\text{UL\_RL}}} = \sum_{r\in T_s^{UL}} r_t\Big(\sum_{i\in L_s^{BS}} \alpha_{irf_s^{\text{UL\_RL}}} - \sum_{n=2}^{U}(n-1)\, z_{f_s^{\text{UL\_RL}}\, rn}\Big) \qquad s\in S \tag*{*(5.49)}$$

$$x_{f_s^{\text{DL\_RL}}} = \sum_{r\in T_s^{DL}} r_t\Big(\sum_{i\in L_s^{BS}} \alpha_{irf_s^{\text{DL\_RL}}} - \sum_{n=2}^{U}(n-1)\, z_{f_s^{\text{DL\_RL}}\, rn}\Big) \qquad s\in S \tag*{*(5.50)}$$

$$\sum_{r\in T_s^{\text{UL}}} \alpha_{irf_s^{\text{UL\_RL}}} = \sum_{r\in T_s^{\text{DL}}} \alpha_{irf_s^{\text{DL\_RL}}} \qquad s\in S,\quad i\in L_s^{BS} \tag*{(5.51)}$$

$$\frac{1+\lambda}{W}\sum_{s\in S}\ \sum_{r\in T_{f_s^{\text{UL\_RL}}}} \left(\frac{\gamma_{sr}^{\text{UL}}\ \nu_s\ b_{ihf_s^{\text{UL\_RL}}}\, \alpha_{irf_s^{\text{UL\_RL}}}}{\dfrac{1}{r_t} + \dfrac{\gamma_{sr}^{\text{UL}}\,\nu_s}{W}} \right. \tag*{*(5.24)}$$

$$\left. - \sum_{n=2}^{U} \frac{\gamma_{sr}^{\text{UL}}\ \nu_s\ b_{ihf_s^{\text{UL\_RL}}}\,(1-SHG_n^{\text{UL}})w_{irf_s^{\text{UL\_RL}}n}}{(1+\dfrac{\gamma_{sr}^{\text{UL}}\,\nu_s\, r_t}{W})\,(\dfrac{1}{r_t}+\dfrac{\gamma_{sr}^{\text{UL}}\,\nu_s\, SHG_n^{\text{UL}}}{W})}\right) \le L_{\text{UL}} \qquad i\in L^{\text{BS}},\quad h\in H$$

$$\sum_{s\in S}\ \sum_{r\in T_{f_s^{\text{DL\_RL}}}} \left(\frac{\gamma_{sr}^{\text{DL}}\ \nu_s \left(N_0 a_{m_s,i} + \frac{P_{\text{BS}}}{W}\big(\sum\limits_{j\in Ring(BS_i)} \frac{a_{m_s,i}}{a_{m_s,j}} + w\big)\right) b_{ihf^{\text{DL\_RL}}\_s}\, \alpha_{irf_s^{\text{DL\_RL}}}}{\dfrac{1}{r_t} + \dfrac{\gamma_{sr}^{\text{DL}}\nu_s w}{W}} \right. \tag*{*(5.25)}$$

$$\left. - \sum_{n=2}^{U} \frac{\gamma_{sr}^{\text{DL}}\ \nu_s \left(N_0 a_{m_s,i} + \frac{P_{\text{BS}}}{W}\big(\sum\limits_{j\in Ring(BS_i)} \frac{a_{m_s,i}}{a_{m_s,j}} + w\big)\right) (1-SHG_n^{\text{DL}})\, b_{ihf_s^{\text{DL\_RL}}}\, w_{irf_s^{\text{DL\_RL}}n}}{(1+\dfrac{\gamma_{sr}^{\text{DL}}\,\nu_s\, w\, r_t}{W})\,(\dfrac{1}{r_t}+\dfrac{\gamma_{sr}^{\text{DL}}\,\nu_s\, w\, SHG_n^{\text{DL}}}{W})}\right)$$

$$\le P_{\text{BS}} - P_{cont} \qquad i\in L^{\text{BS}},\quad h\in H$$

Wired Link Capacity

$$*(5.23)\qquad \sum_{s\in S}\sum_{v\in V_{f_s^{\text{UL\_WL}}}} a_{hs}\, m_{vl}\, x^{v}_{f_s^{\text{UL\_WL}}} + \sum_{s\in S}\sum_{v\in V_{f_s^{\text{DL\_WL}}}} a_{hs}\, m_{vl}\, x^{v}_{f_s^{\text{DL\_WL}}} \le C_\ell$$

$$h \in H, \ell = 1, 2, ..., m$$

$$(5.22)\qquad \sum_{s\in S}\sum_{f\in F_s^{\text{WL}}} b_{ihf} x^{i}_{f} \le C_{\ell_{BS_i}} \qquad h \in H, i \in L^{\text{BS}}$$

$$(5.12)\qquad x^{i}_{f_s^{\text{DL\_WL}}} = \rho_{f_s^{\text{DL\_WL}}} \sum_{r\in T_s^{DL}} \alpha_{irf_s^{\text{DL\_RL}}} \qquad s \in S^{CBR}, \quad i \in L_s^{BS}$$

$$(5.13)\qquad x^{i}_{f_s^{\text{UL\_WL}}} = \rho_{f_s^{\text{UL\_WL}}} \sum_{r\in T_s^{UL}} \alpha_{irf_s^{\text{UL\_RL}}} \qquad s \in S^{CBR}, \quad i \in L_s^{BS}$$

$$*(5.14)\qquad \sum_{r\in T_s^{UL}} r_t \alpha_{irf_s^{\text{UL\_RL}}} \le x^{i}_{f_s^{\text{UL\_WL}}} \le K_1 \sum_{r\in T_s^{UL}} r_t \alpha_{irf_s^{\text{UL\_RL}}} \quad s \in S^{VBR}, \quad i \in L_s^{BS}$$

$$*(5.15)\qquad 0 \le x^{i}_{f_s^{\text{DL\_WL}}} \le \sum_{r\in T_s^{UL}} r_t \alpha_{irf_s^{\text{DL\_RL}}} \quad s \in S^{VBR}, \qquad\qquad i \in L_s^{BS}$$

$$*(5.29)\qquad x^{i}_{f_s^{\text{UL\_WL}}} \le \sum_{v\in V^{k}_{f_s^{\text{UL\_WL}}}} x^{v}_{f_s^{\text{UL\_WL}}} \le \sum_{j\in L_s^{BS}\cap N_{BS_k}} x^{j}_{f_s^{\text{UL\_WL}}}$$

$$s \in S, \quad k \in K_s, \quad i \in L_s^{BS} \cap N_{BS_k}$$

$$*(5.30)\qquad x^{i}_{f_s^{\text{DL\_WL}}} \le \sum_{v\in V^{k}_{f_s^{\text{DL\_WL}}}} x^{v}_{f_s^{\text{DL\_WL}}} \le \sum_{j\in L_s^{BS}\cap N_{BS_k}} x^{j}_{f_s^{\text{DL\_WL}}}$$

$$s \in S, \quad k \in K_s, \quad i \in L_s^{BS} \cap N_{BS_k}$$

Soft handoff

$$*(5.31) \qquad \sum_{n=2}^{U} \sum_{r \in T_s^{\text{UL}}} z_{f_s^{\text{UL\_RL}} rn} \leq y_s \qquad s \in S$$

$$*(5.32) \qquad \sum_{n=2}^{U} \sum_{r \in T_s^{\text{DL}}} z_{f_s^{\text{DL\_RL}} rn} \leq y_s \qquad s \in S$$

$$*(5.35) \qquad y_s = \sum_{i \in L_s^{\text{BS}}} \sum_{r \in T_s^{\text{UL}}} \alpha_{irf_s^{\text{UL\_RL}}} - \sum_{n=2}^{U} \sum_{r \in T_s^{\text{UL}}} (n-1)\, z_{f_s^{\text{UL\_RL}} rn} \qquad s \in S$$

$$*(5.36) \qquad y_s = \sum_{i \in L_s^{\text{BS}}} \sum_{t \in T_s^{\text{DL}}} \alpha_{irf_s^{\text{DL\_RL}}} - \sum_{n=2}^{U} \sum_{r \in T_s^{\text{DL}}} (n-1)\, z_{f_s^{\text{DL\_RL}} rn} \qquad s \in S$$

$$*(5.38) \qquad z_{f_s^{\text{UL\_RL}} rn} = \frac{1}{n} \sum_{i \in L_s^{BS-SH}} w_{irf_s^{\text{UL\_RL}} n} \qquad r \in T_s^{\text{UL}}, \quad s \in S, \quad n = 2,..,U$$

$$*(5.39) \qquad z_{f_s^{\text{DL\_RL}} rn} = \frac{1}{n} \sum_{i \in L_s^{BS-SH}} w_{irf_s^{\text{DL\_RL}} n} \qquad r \in T_s^{\text{DL}}, \quad s \in S, \quad n = 2,..,U$$

$$*(5.40) \qquad \alpha_{irf_s^{\text{UL\_RL}}} \geq \sum_{n=2}^{U} w_{irf_s^{\text{UL\_RL}} n} \qquad r \in T_s^{\text{UL}}, \quad s \in S, \quad i \in L_s^{BS-SH}$$

$$*(5.41) \qquad \alpha_{irf_s^{\text{DL\_RL}}} \geq \sum_{n=2}^{U} w_{irf_s^{\text{DL\_RL}} n} \qquad r \in T_s^{\text{DL}}, \quad s \in S, \quad i \in L_s^{BS-SH}$$

$$*(5.47) \qquad \sum_{n=2}^{U} \sum_{r \in T_s^{\text{UL}}} (w_{irf_s^{\text{UL\_RL}} n} + w_{jrf_s^{\text{UL\_RL}} n}) \leq 1$$

$$s \in S, \quad r \in T_s^{\text{UL}}, \quad n = 2,..,U, \quad i,j \in L_s^{BS-SH}, BSC_i \neq BSC_j$$

$$*(5.48) \qquad \sum_{n=2}^{U} \sum_{r \in T_s^{\text{DL}}} (w_{irf_s^{\text{DL\_RL}} n} + w_{jrf_s^{\text{DL\_RL}} n}) \leq 1$$

$$s \in S, \quad r \in T_s^{\text{DL}}, \quad n = 2,..,U, \quad i,j \in L_s^{BS-SH}, BSC_i \neq BSC_j$$

Delay Constraints

$$* (5.21) \quad x^{i}_{f_s^{\text{DL\_WL}}} \geq$$

$$\sum_{r \in T_s^{DL}} \max(\mathcal{R}(r_t), \frac{m^{max}_{f_{\text{DL}}} L^{max}_{f\text{DL}} + \sigma_s}{D_S^{\text{DL\_MAX}} - D^{max-transe,rt}_{f_s^{\text{DL\_WL}}} - D^{max-propag}_{f_s^{\text{DL\_WL}}} - D^{max-trait}_{f_s^{\text{DL\_WL}}} - D^{r}_{f_s^{\text{DL\_RL}}}}) \alpha_{ir f_s^{\text{DL\_RL}}}$$

$$s \in S^{\text{VBR}}, \quad i \in L_s^{\text{BS}}$$

Selection of Active Base Stations

$$*(5.54) \qquad y_i^{\text{BS}} \leq \sum_{s \in S} \sum_{r \in T_{f_s^{\text{DL\_RL}}}} \alpha_{irf} \qquad i \in L^{\text{BS}}$$

$$*(5.55) \qquad \sum_{r \in T_{f_s^{\text{DL\_RL}}}} \alpha_{irf} \leq y_i^{\text{BS}} \qquad i \in L_s^{\text{BS}}, \qquad s \in \text{S}$$

5.7 Improved model

The proposed model in [1] has been modified in [40]. The modifications are based on the following observation: there always exists an optimal solution of the model in [1] in which the value of bandwidth on the wired links are the smallest possible. Note that this is not necessarily the case for all optimal solutions since a session that is active during a less loaded period may use a greater bandwidth while still satisfying the capacities of the links. Considering the first case where we have the optimal solution with the smallest possible bandwidth, we are able to replace some inequality constraints by equalities. Consequently, the feasibility domain becomes tighter, which results in a better lower bound for continuous relaxation. This helps in the pruning of the non-promising branches. Therefore, it gives us the opportunity to obtain feasible solutions with values that are much closer to the optimal one, i.e. with a better precision.

Based on the same reasoning we have changed the model proposed in this chapter.

The fact that the potential base stations which can serve a session may be connected to more than one BSC complicates these changes comparing to [40]. Since the objective includes minimizing the capacity on the wired links of core network, in the case of VBR sessions in which the value of bandwidth for the flow f on the wired links lays between a lower bound and an upper bound we can assign the minimum value of bandwidth, $\underline{\rho}_f$, on the wired links in the core network. For instance, in the uplink direction where the bandwidth on the radio link should be smaller than the bandwidth on wired link we choose the maximum value between the data rate of the chosen RAB and the minimum possible value of bandwidth on wired links.

Since in this mathematical model the number of variables and constraints are increasing compared to [1], these changes are crucial for running the program with big instances in terms of the number of requested sessions, number of base stations and number of possible paths for each session.

For the CBR sessions, where $x_f^{\text{BSC}_k} = \sum\limits_{v \in V_f^k} x_f^v$ and $\alpha_{rf} = \sum_{i \in L_s^{\text{BS}}} \alpha_{irf} - \sum\limits_{n=2}^{U} (n-1)\, z_{frn}$ the new constraints are:

$$\sum_{k \in K_s} x_{f_s^{\text{UL\_WL}}}^{\text{BSC}_k} = \rho_{f_s^{\text{UL\_WL}}} \sum_{r \in T_{f_s^{\text{UL\_RL}}}} \alpha_{r f_s^{\text{UL\_RL}}} \qquad s \in S^{\text{CBR}}, \tag{5.56}$$

$$\sum_{k \in K_s} x_{f_s^{\text{DL\_WL}}}^{\text{BSC}_k} = \rho_{f_s^{\text{DL\_WL}}} \sum_{r \in T_{f_s^{\text{DL\_RL}}}} \alpha_{r f_s^{\text{DL\_RL}}} \qquad s \in S^{\text{CBR}}. \tag{5.57}$$

α_{rf} is a binary variable which is equal to one if RAB r is selected. When the session is CBR then for each flow f, there is a declared amount of bandwidth on the wired links called ρ_f. In constraints (5.56) and (5.57) for the session s if the RAB r is selected the bandwidth on the path which goes through the BSC k is equal to $\rho_{f_s^{\text{UL\_WL}}}$ and $\rho_{f_s^{\text{DL\_WL}}}$ in uplink and downlink directions respectively.

For the VBR sessions, by considering the following equations

$$x_f^{\text{BSC}_k} = \sum_{v \in V_f^k} x_f^v,$$

$$\alpha_{rf} = \sum_{i \in L_s^{\text{BS}}} \alpha_{irf} - \sum_{n=2}^{U} (n-1)\, z_{frn},$$

$$\underline{\rho}_s^{\text{DL}}(r) = \max\left(\mathcal{R}(r_t), \frac{m_{f_{\text{DL}}}^{max}\, L_{f\text{DL}}^{\max} + \sigma_s}{D_s^{\text{DL\_MAX}} - D_{f_s^{\text{DL\_WL}}}^{max-transe,rt} - D_{f_s^{\text{DL\_WL}}}^{max-propag} - D_{f_s^{\text{DL\_WL}}}^{max-trait} - D_{f_s^{\text{DL\_RL}}}^{r}}\right),$$

the new constraints are:

$$x^i_{f_s^{\text{UL\_WL}}} = \sum_{r \in T_{f_s^{\text{UL\_RL}}}} \max(r_t, \underline{\rho}_{f_s^{\text{UL\_WL}}})\alpha_{irf_s^{\text{UL\_RL}}} \qquad i \in L_s^{\text{BS}}, s \in S^{\text{VBR}}, \qquad (5.58)$$

$$x^i_{f_s^{\text{DL\_WL}}} = \sum_{r \in T_{f_s^{\text{DL\_RL}}}} \max(\underline{\rho}_s^{\text{DL}}(r), \underline{\rho}_{f_s^{\text{DL\_WL}}})\alpha_{irf_s^{\text{DL\_RL}}} \qquad i \in L_s^{\text{BS}}, s \in S^{\text{VBR}}. \qquad (5.59)$$

When session s is VBR for each flow f there are a minimum and a maximum of bandwidth on the wired links: $\underline{\rho}$, $\overline{\rho}$. In the uplink direction if the session is served by base station i and r is the selected RAB then $x^i_{f_s^{\text{UL\_WL}}}$ takes the maximum value between the data rate, r_t, of RAB r and the minimum bandwidth for the flow f on the wired links. In the downlink direction, however, referring to Section 4.2.1 the bandwidth on the wired links should be smaller than the bandwidth on the radio links to prevent the accumulation on the base stations level. Hence, we introduce $\mathcal{R}(r_t)$, see Table 4.3. Therefore for the session s if RAB r is chosen, then $x^i_{f_s^{\text{DL\_WL}}}$ will be equal to the maximum value between $\underline{\rho}_s^{\text{DL}}(r)$ and the minimum bandwidth for the flow f on the wired links.

$$\sum_{k \in K_s} x^{\text{BSC}_k}_{f_s^{\text{UL\_WL}}} = \sum_{r \in T_{f_s^{\text{UL\_RL}}}} \max(r_t, \underline{\rho}_{f_s^{\text{UL\_WL}}})\alpha_{rf_s^{\text{UL\_RL}}} \qquad s \in S^{\text{VBR}}, \qquad (5.60)$$

$$\sum_{k \in K_s} x^{\text{BSC}_k}_{f_s^{\text{DL\_WL}}} = \sum_{r \in T_{f_s^{\text{DL\_RL}}}} \max(\underline{\rho}_s^{\text{DL}}(r), \underline{\rho}_{f_s^{\text{DL\_WL}}})\alpha_{rf_s^{\text{DL\_RL}}} \qquad s \in S^{\text{VBR}}. \qquad (5.61)$$

Having these new constraints in hand, we can remove constraints (5.14) and (5.15). Constraint (5.21) is dominated by constraint (5.59), therefore it can be removed as well.

5.7.1 Improved dimensioning mathematical model: Summing Up

We sum up below the set of constraints of the anticipative mathematical model. Only the modified constraints are displayed.

$$\min \left(w_c \sum_{\ell=1}^{m} d_\ell \, C_\ell + w_c \sum_{i=1}^{n_{\text{BS}}} d_{\ell_{\text{BS}i}} C_{\ell_{\text{BS}i}} + w_{BS} \sum_{i=1}^{n_{\text{BS}}} y_i^{\text{BS}} \right)$$

Call Admission Control:

When the session is CBR, no change: (5.2), (5.3), (5.4), (5.5).

When the session is VBR, no change: (5.6), (5.7), (5.8), (5.9), and also (5.10), (5.11).

Grade of Service:

No change: (5.26).

Quality of Service:

No change: (5.52), (5.53).

Radio Link Capacity:

No change: (5.49), (5.50), (5.51), (5.24), (5.25).

Wired Link Capacity:

$$\sum_{s\in S}\sum_{v\in V_{f_s^{\mathrm{UL\_WL}}}} a_{hs}\, m_{vl}\, x^{v}_{f_s^{\mathrm{UL\_WL}}} + \sum_{s\in S}\sum_{v\in V_{f_s^{\mathrm{DL\_WL}}}} a_{hs}\, m_{vl}\, x^{v}_{f_s^{\mathrm{DL\_WL}}} \le C_\ell \qquad h\in H, \ell = 1,2,\ldots,m \tag{5.23}$$

$$\sum_{s\in S}\sum_{f\in F_s^{\mathrm{RL}}} b_{ihf} x_f^i \le C_{\ell_{BS_i}} \qquad h\in H, i\in L^{\mathrm{BS}} \tag{5.22}$$

$$x^i_{f_s^{\mathrm{DL\_WL}}} = \rho_{f_s^{\mathrm{DL\_WL}}} \sum_{r\in T_s^{DL}} \alpha_{ir f_s^{\mathrm{DL\_RL}}} \qquad s\in S^{CBR}, \quad i\in L_s^{BS} \tag{5.12}$$

$$x^i_{f_s^{\mathrm{UL\_WL}}} = \rho_{f_s^{\mathrm{UL\_WL}}} \sum_{r\in T_s^{UL}} \alpha_{ir f_s^{\mathrm{UL\_RL}}} \qquad s\in S^{CBR}, \quad i\in L_s^{BS} \tag{5.13}$$

$$\sum_{k\in K_s} x^{\mathrm{BSC}_k}_{f_s^{\mathrm{UL\_WL}}} = \rho_{f_s^{\mathrm{UL\_WL}}} \sum_{r\in T_{f_s^{\mathrm{UL\_RL}}}} \alpha_{r f_s^{\mathrm{UL\_RL}}} \qquad s\in S^{\mathrm{CBR}} \tag{5.56}$$

$$\sum_{k\in K_s} x^{\mathrm{BSC}_k}_{f_s^{\mathrm{DL\_WL}}} = \rho_{f_s^{\mathrm{DL\_WL}}} \sum_{r\in T_{f_s^{\mathrm{DL\_RL}}}} \alpha_{r f_s^{\mathrm{DL\_RL}}} \qquad s\in S^{\mathrm{CBR}} \tag{5.57}$$

$$x^i_{f_s^{\mathrm{UL\_WL}}} = \sum_{r\in T_{f_s^{\mathrm{UL\_RL}}}} \max(r_t, \underline{\rho}_{f_s^{\mathrm{UL\_WL}}})\alpha_{ir f_s^{\mathrm{UL\_RL}}} \qquad i\in L_s^{\mathrm{BS}}, s\in S^{\mathrm{VBR}} \tag{5.58}$$

$$x^i_{f_s^{\mathrm{DL\_WL}}} = \sum_{r\in T_{f_s^{\mathrm{DL\_RL}}}} \max(\underline{\rho}_s^{\mathrm{DL}}(r), \underline{\rho}_{f_s^{\mathrm{DL\_WL}}})\alpha_{ir f_s^{\mathrm{DL\_RL}}} \qquad i\in L_s^{\mathrm{BS}}, s\in S^{\mathrm{VBR}} \tag{5.59}$$

$$\sum_{k\in K_s} x^{\mathrm{BSC}_k}_{f_s^{\mathrm{UL\_WL}}} = \sum_{r\in T_{f_s^{\mathrm{UL\_RL}}}} \max(r_t, \underline{\rho}_{f_s^{\mathrm{UL\_WL}}})\alpha_{r f_s^{\mathrm{UL\_RL}}} \qquad s\in S^{\mathrm{VBR}} \tag{5.60}$$

$$\sum_{k\in K_s} x^{\mathrm{BSC}_k}_{f_s^{\mathrm{DL\_WL}}} = \sum_{r\in T_{f_s^{\mathrm{DL\_RL}}}} \max(\underline{\rho}_s^{\mathrm{DL}}(r), \underline{\rho}_{f_s^{\mathrm{DL\_WL}}})\alpha_{r f_s^{\mathrm{DL\_RL}}} \qquad s\in S^{\mathrm{VBR}} \tag{5.61}$$

$$x^i_{f_s^{\mathrm{UL\_WL}}} \le \sum_{v\in V^k_{f_s^{\mathrm{UL\_WL}}}} x^v_{f_s^{\mathrm{UL\_WL}}} \le \sum_{j\in L_s^{BS}\cap N_{BS_k}} x^j_{f_s^{\mathrm{UL\_WL}}} \qquad s\in S, \quad k\in K_s, \quad i\in L_s^{BS}\cap N_{BS_k} \tag{5.29}$$

$$x^i_{f_s^{\mathrm{DL\_WL}}} \le \sum_{v\in V^k_{f_s^{\mathrm{DL\_WL}}}} x^v_{f_s^{\mathrm{DL\_WL}}} \le \sum_{j\in L_s^{BS}\cap N_{BS_k}} x^j_{f_s^{\mathrm{DL\_WL}}} \qquad s\in S, \quad k\in K_s, \quad i\in L_s^{BS}\cap N_{BS_k} \tag{5.30}$$

Soft handoff:

No change: (5.31), (5.32), (5.35), (5.36), (5.38), (5.39), (5.40), (5.41), (5.47), (5.48).

Selection of Active Base Stations:

No change: (5.54), (5.55).

Chapter 6

Implementation and Results

The optimization model described in the previous chapters has been implemented in C++, using the libraries of CPLEX-MIP [41] in order to solve the corresponding mixed linear program. The implementation has two modules:

- The network and traffic problem generator,
- The variables and constraints of the MIP program.

6.1 Instance generator

The problem generator generates all necessary traffic files for a given problem instance. The traffic files will be used in order to generate the variables and constraints of a MIP problem, which will be solved by CPLEX-MIP.

There are some input data as well as a network architecture, which are needed to generate the traffic files. Essentially the generator first places the base stations uniformly in a given surface, then places the requested sessions and assigns their beginning and duration based on a fixed simulation time, as well as their priority, list of possible paths and their list of potential base stations.

6.1.1 Network architecture

As mentioned above the network architecture is given for each instance. This file contains the number of BSCs, external nodes and all possible paths between each BSC and external node. This information does not change during the simulation time and is mostly used for defining the routing table and set of possible paths for each individual session.

6.1.2 Simulation space, base stations, base stations in soft handoff and paths

The length and width of the simulation space are given as input data:

- LONGUEUR_SURFACE,
- LARGEUR_SURFACE.

The number of BSCs can be extracted from the network architecture, Section 6.1.1, then the simulation space is divided equally between the existed BSCs. In this model no already existed base station is considered in the surface and we place all the base stations on this plain surface uniformly. This assumption is important for the capacity formulas in radio links, Section 4.1.4. The number of base stations is again known from the beginning of the simulation and is an input data named:

- NBR_STATIONS_POTENTIELLES.

In this distribution each base station is connected to only one BSC, and that is the one which covers the geographical position of the base station on the simulation surface. In this context the soft handoff is possible between two or more base stations. In each simulation we specify the number of base stations involved in soft handoff cases for each session from the beginning of the simulation.

- ➤ NBR_BS_in_SHO.

Between each pair of origin destination nodes, there may be more that one possible path which can serve a session. For one session the number of possible paths may be equal to 4 and for another may be 5. In this thesis for each simulation, a limited number of paths in multi-routing has been considered.

- ➤ NBR_MULTI_PATH.

Assume that we have considered $NBR\_MULTI\_PATH = 3$. For each session, if the number of possible paths set is more than 3 then only the first three paths in the set of possible paths will be taken into account. Obviously, when the number of possible paths set for a session is less than 3, for example is equal to 2, both paths will be considered.

6.1.3 Initial solution

There is the possibility to solve a problem instance starting from a given initial solution. We can also check the feasibility of a solution obtained by another program. In order to check this we consider that solution as an initial solution. The feasibility of the initial solution is first checked and if the result is positive the optimization procedure goes on with that initial solution.

6.1.4 Simulation planning time

The simulation planning time is an input data and is divided in periods. The number of periods can be different for each traffic instance. The parameters are:

- ➤ DUREE_PLANIFICATION,
- ➤ NBR_PERIODE.

As seen in Section 4.2.1 the beginning and the finish time of each session coincide with those of the periods. Note that the number of periods play an important role in the simulation. Obviously, it will change the length of the periods. The more the number of the periods are, the shorter the length of the periods become. In addition the more the number of periods are, the less sessions are in the same period. This reduces the resource reservation on each period. On the other hand, the number of the periods has an impact on the number of constraints. More precisely, it has an effect on the following constraints: (5.24), (5.25), (5.23) and (5.22). Therefore we cannot just increase the numbers of periods as much as we want, since it influences the size of problem and its solution time.

6.1.5 Session generation

A session is a connection between two entities and can be of voice, video streaming, videophone, web browsing or mail type, see Table 6.1. Each of these applications belongs to one of the fourth class of applications discussed in Chapter 3. The total number of sessions is known from the beginning of each simulation and is considered as an input data:

- ➤ NBR_TOTAL_SESSIONS.

For each instance we define the proportion of each type of application. Referring to [42] we consider two scenarios named scenario A and scenario B. Scenario A considers a high utilization of multi-media applications while scenario B supports the multi-media applications as well, but gives a priority to the voice application over the other ones. We use [42] for the distribution of the sessions among the various applications in each scenario. Table 6.2 demonstrates the percentage of users per application in both scenarios.

Scenario is another input data for each traffic instance called:

➤ SCENARIO

At this point, since we know the total number of sessions and the selected scenario, the model can conclude for the number of sessions for each type of application. Once the number of sessions per application is determined, for each session we define the geographical position, beginning, duration, priority, type (High and Low for video conference, videophone and web browsing), direction for mail (the other types are in both directions and video streaming is only in downlink), set of possible paths and set of potential base stations, Table 6.1. For each session we also define a mother-BSC. This is the BSC which covers the area in which the session is located. However, a session can be served by a base station connected to a BSC other than its mother-BSC.

	Voice	**Videophone**	**Video Streaming**	**Web Browsing**	**Mail**
	% of user	% of user	% of user	% of user	% of user
Priority	√	√	√	√	√
Type (High/Low)	-	√	√	√	-
Start of the session	√	√	√	√	√
Length of the session	√	√	√	√	√
Position	√	√	√	√	√
Direction (UL/DL)	-	-	DL	-	DL or UL
Set of potential BSs	√	√	√	√	√
Set of potential paths	√	√	√	√	√

TAB. 6.1: Session specifications based on its type of application

Scenario	**Voice**	**Videophone**	**Video Streaming**	**Web Browsing**	**Mail**
	% of user	% of user	% of user	% of user	% of user
A	40%	5%	10%	25%	20%
B	60%	3%	8%	12%	17%

TAB. 6.2: Application utilization by the users

In order to specify the priority, type and direction of a session the Bernoulli laws are used. There are parameters which define the probabilities of generating a Gold session, a High session and a session in the downlink direction. For instance, if *param* denotes the probability of the uplink direction for a mail session then *(1-param)* defines the probability of having a mail session in downlink direction. The choice of direction is important for the mail sessions, since mail sessions are either uplink or downlink, but not both.

The session interval times are generated by two distribution laws, exponential for voice, video streaming and mail applications and lognormal distribution for videophone and web browsing. Then, the beginning of each session is reset to the beginning of the period during which the session has been started.

- Exponential distribution describes the distance between the sessions with uniform distribution in time. The arrival times are based on a Poisson distribution. λ is the parameter for time variable and $\frac{1}{\lambda}$ and $\frac{1}{\lambda^2}$ are the average of interval times and the variance.

- In lognormal distribution the parameters are the average and the variance as in the exponential distribution, $\frac{1}{\lambda}$ for average of intervals and $\frac{1}{\lambda^2}$ for the variance.

In both distributions $\lambda = \frac{\text{number of sessions for an application type}}{\text{planning time}}$. Therefore the arrival rate is fixed during the planning time. In order to calculate the duration of a session we consider an average duration for each type of applications and we use the exponential distribution with parameter $\lambda = \frac{1}{\text{average duration of the application type}}$. The numerical values and the parameters used in the program are specified in Tables 6.3, 6.4, 6.5, 6.6 and 6.7 for each type of applications, we use the same than in [1].

Voice	Distribution	Parameter	Value of the parameter
Arrival time	Poisson	λ	$\frac{\text{number of voice sessions}}{\text{planning time}}$
Interval times	exponential	λ	$\frac{\text{number of voice sessions}}{\text{planning time}}$
Duration	exponential	Average duration	31 sec
Priority	Bernoulli	Gold probability	0.7

TAB. 6.3: Parameter values for the voice application

Videophone	Distribution	Parameter	Value of the parameter
Interval times	lognormal	$average = \frac{1}{\lambda}$ $variance = \frac{1}{\lambda^2}$	$\lambda = \frac{\text{number of videophone sessions}}{\text{planning time}}$
Duration	exponential	Average duration	180 sec
Priority	Bernoulli	Gold probability	0.7
Type	Bernoulli	High probability	0.3

TAB. 6.4: Parameter values for the videophone application

Video streaming	Distribution	Parameter	Value of the parameter
Arrival time	Poisson	λ	$\frac{\text{number of video streaming sessions}}{\text{planning time}}$
Interval times	exponential	λ	$\frac{\text{number of video streaming sessions}}{\text{planning time}}$
Duration	exponential	Average duration	180 sec
Priority	Bernoulli	Gold probability	0.7
Type	Bernoulli	High probability	0.3

TAB. 6.5: Parameter values for the video streaming application

Web browsing	Distribution	Parameter	Value of the parameter
Interval times	lognormal	$average = \frac{1}{\lambda}$ $variance = \frac{1}{\lambda^2}$	$\lambda = \frac{\text{number of web browsing sessions}}{\text{planning time}}$
Duration	exponential	Average duration	250 sec
Priority	Bernoulli	Gold probability	0.5
Type	Bernoulli	High probability	0.3

TAB. 6.6: Parameter values for the web browsing application

6.1.6 Session position

The geographical positions of the sessions are important in order to specify the mother-BSC, potential base stations which can give service to the session and consequently the set of possible paths. The generation of the sessions are homogeneous on all the periods, since the mobile stations which ask for the same application type at the same period are placed uniformly on the simulation surface.

Mail	Distribution	Parameter	Value of the parameter
Arrival time	Poisson	λ	$\frac{\text{number of video mail sessions}}{\text{planning time}}$
Interval times	exponential	λ	$\frac{\text{number of mail sessions}}{\text{planning time}}$
Duration	exponential	Average duration	10 sec
Priority	Bernoulli	Gold probability	0.5
Direction	Bernoulli	Downlink probability	0.5

TAB. 6.7: Parameter values for the mail application

6.1.7 Choice of base stations

Once the geographical position of each session is determined, the problem generator defines a set of potential base stations for each session. We call this set L_s^{BS} for session s. The base stations which belong to this set can serve the session s both is soft handoff and non soft handoff. We divide this set in two subsets. One subset is $L_s^{\text{BS\_NSH}}$ which cannot serve session s in soft handoff and the other is $L_s^{\text{BS\_SH}}$ which can serve the session s both in soft handoff and non soft handoff. Therefore we have:

$$L_s^{\text{BS}} = L_s^{\text{BS\_NSH}} + L_s^{\text{BS\_SH}} . \tag{6.1}$$

The choice of the base stations is based on the distance between each base station and the session. Two distances are defined as input data:

- DISTANCE_MAXIMUM_SHO d_{sho},
- DISTANCE_MAXIMUM_NON_SHO d_{nsh}.

The base stations which are located in a circle with radius equal to d_{nsh} around a session, can serve that session in non soft handoff. These are the base stations which are relatively close to the session, while the base stations which are located in a circle with radius equal to d_{sho} around a session, can serve that session both in non soft handoff and soft handoff. Figure 6.1 illustrates the base stations choice for each session.

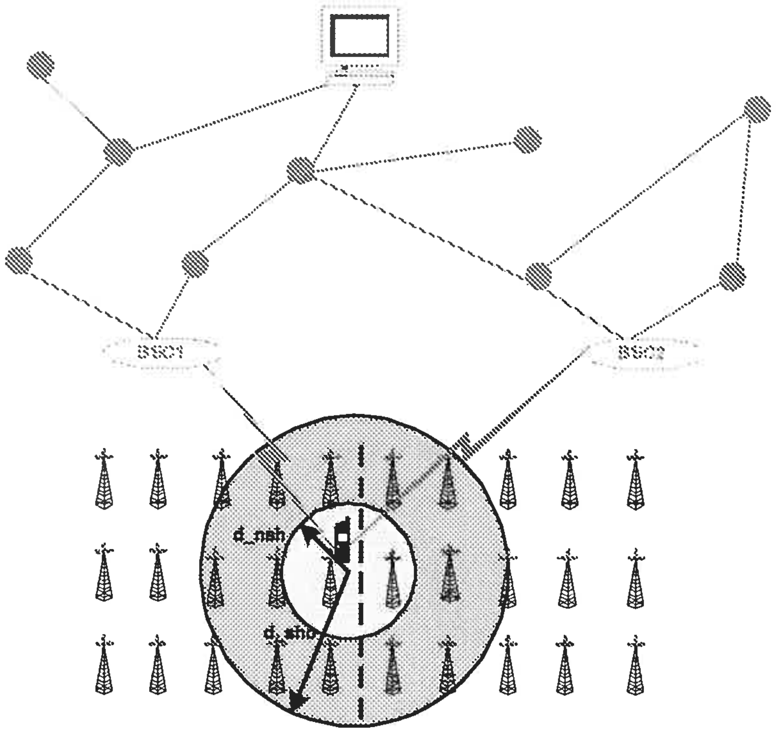


FIG. 6.1: Set of potential base stations for each session

The selected base stations are not necessarily connected to the same BSC. They may be connected to different BSCs depending on the geographical position of the session. In the input files the d_{sho} is fixed to 150 m while the d_{nsh} is fixed to 50 m. We make sure that there is a reasonable number of potential base stations for each session, while choosing these distances.

For each base station which belongs to L_s^{BS}, the attenuation factor coefficient will be calculated as well. This coefficient is used in the radio capacity formula in downlink direction, constraint (5.25). Recalling the formula from Section 4.1.4 we have:

$$a_{s,i} = \frac{d^4}{h_i^2 h_s^2} \tag{6.2}$$

where $a_{s,i}$ is the attenuation factor between the base station i with height $h_i = 30$ meters and the session s with height $h_s = 2$ meters. d represents the distance between the base station i and the session s.

6.1.8 Choice of paths

For each session, the problem generator defines a set of possible paths from its source to its destination. The destination is an external node and is randomly chosen for each session individually. The choice of paths is related to the set of potential base stations. If all the potential base stations are connected to one BSC, i.e. the mother-BSC, then only the paths which go through that BSC to the selected external node will be considered, see Figure 4.5. On the other hand, if there are potential base stations that are connected to different BSCs other than the mother-BSC, then we have two types of paths. First the paths which go through those BSCs and reach the selected external node and second the paths which pass the mother-BSC, see Figure 4.6. Note that the first path in the set of potential routes will be always a path which go through the mother-BSC of the session. This restriction has been considered for the cases where we run the program with only one possible path. The complete description of paths choice can be found in Section 4.3. The number of potential paths in a set for each session has an impact on the size of the problem as it directly influences the number of binary and continuous variables. The greater the number of potential paths the more binary and continuous variables we will have. Consequently, the problem will be much harder to be solved. As mentioned before for each instance we set a limit on the number of paths in multi-routing. While comparing multi and mono-routing we have considered the value 3 for the NBR_MULTI_PATH.

6.2 Variables and constraints of the MIP program

The problem generator uses the traffic files built by the traffic generator, based on input data, and creates the constraints and the objective function of the proposed dimensioning problem. The constraints are generated in a format that can be read by the CPLEX optimization software. At this point the CPLEX-MIP tries to solve the problem.

The solution generated by CPLEX-MIP, *solution.txt*, which lets us analyze each session in terms of the bandwidth, RAB, selected base station, or base stations in case of soft handoff, and routing paths for uplink and downlink directions. A problem can also be solved using an initial solution. The values for all variables (binary and continuous) is then taken from the *solution.txt* file and CPLEX-MIP tries to find a better solution. This helps us to gain better solutions while starting from an initial solution and a less amount for gap parameter in CPLEX [41].

In this section we discuss the value of parameters used in generating the objective function and the constraints.

6.2.1 Parameter values in objective function

In this project we aim at minimizing the number of activated base stations in the radio network and used capacity on the wired links both in core and radio networks. The objective function is discussed in Section 5.3. The value for each parameter is:

- $w_{BS} = 500$,
- $w_c = 1$,
- $d_l = 1$.

When we talk about the costs for the base stations and wired links we mean both the cost of installation-maintenance and the cost of buying the equipments.

6.2.2 Values for bandwidth of the flow

As mentioned in Section 2.8.4 for each session the value of bandwidth is fixed if it is a CBR session and lies between an upper bound and lower bound in the case of VBR session. The values used in this model are described in Tables 6.8, 6.9, 6.10 and 6.11.

ρ_s kbps	Voice	Videophone High	Videophone Low	Video Streaming High	Video Streaming Low
UL_RL	8	144	64	-	-
DL_RL	8	144	64	144	32

TAB. 6.8: Bandwidth on radio links for CBR applications

t_s kbps	Voice	Videophone High	Videophone Low	Video Streaming High	Video Streaming Low
UL_RL	9.6	153.6	76.8	-	-
DL_RL	9.6	153.6	76.8	153.6	38.4

TAB. 6.9: Bandwidth on wired links for CBR applications

6.2.3 Set of RABs, values for ratio of signal to noise $\gamma_{s,r}$ and Quality of Service coefficient $Q_{r,a,p}$ per application type

The values for data rate and Frame Error Rate associated to each RAB r, as well as the ratio of signal energy per bit to noise spectral density $\gamma_{s,r}$, Section 4.1.4, are taken from [36], where two different types of users are considered: a pedestrian with channel speed of 3 km/hr and a vehicular with 30 km/hr channel speed.

The users are considered to be static. Therefore only the values associated with a pedestrian user in [36] have been used in our model.

The proposition of RAB for the web browsing and mail applications are based on making the delay constraint realistic. This is not an issue for the other applications, since the delay constraint is applicable just on VBR applications, Section 4.4.1.

6.2.4 Numerical values for constant parameters in radio capacity formulas

There are several constants used in the radio capacity formulas and their definition can be found in Section 4.1.4. The numerical values for the constants are taken from [5]. Note that [5] has studied WCDMA for UMTS networks, therefore the value of

$\underline{\rho}_s/\overline{\rho}_s$	**Web browsing**	**Web browsing**	**Mail**
kbps	High	Low	
UL_RL	10/25	8/15	10/300
DL_RL	100/400	20/150	10/300

TAB. 6.10: Bandwidth on radio links for VBR applications

$\underline{t}_s/\overline{t}_s$	**Web browsing**	**Web browsing**	**Mail**
kbps	High	Low	
UL_RL	9.6/307.2	9.6/307.2	9.6/153.6
DL_RL	9.6/307.2	9.6/307.2	9.6/153.6

TAB. 6.11: Bandwidth on wired links for VBR applications

bandwidth used in our experiments has been changed to the bandwidth for CDMA2000. The ratio of $\sum_{i \in Ring(BS_0)} \frac{a_{s,0}}{a_{s,i}}$ depends on the six base stations around the BS_0. These base stations are not known at the beginning. Therefore the value of $\sum_{i \in Ring(BS_0)} \frac{a_{s,0}}{a_{s,i}}$ cannot be known either. In [35] the average ratio of $E[\frac{a_{s,0}}{a_{s,i}}]$ is equal to 0.117. We take the same value and consider $0.117 = \frac{a_0}{a_{s,j}}$ where a_0 is the attenuation factor for a mobile station which its serving base station is placed around the cell's radius far from the mobile station. As the same value should be considered for all the six neighbor base

Parameter	**Recommended value**
W	1.25 Mhz
λ	55%
v_s	0.67 for voice sessions (speech) 1.0 for other sessions (data)
P_{BS}	20 Watts
P_{cont}	4 Watts = 20%P_{BS}
w	0.4
SHG_{UL}	45% which corresponds to a gain of 3.5 db
SHG_{DL}	45%
L_{UL}	60%
N_o	1.258910^{-20} W/Hz

TAB. 6.12: Parameter values for capacity formulas

stations around the BS_0 we have:

$$\sum_{i \in Ring(BS_0)} \frac{a_{s,0}}{a_{s,j}} = 6 \times a_{s,i} \times \frac{0.117}{a_0}.$$

Since the $N_0 \times a_{s,i}$ is negligible in the downlink radio capacity formula, this approximation is taken into account in order to keep the relation between the radio capacity in the downlink and the position of sessions. The attenuation factor, $a_{s,i}$ between the mobile station i and the base station s can be calculated based on the distance between them. The distance is also a known value from the beginning which is calculated in traffic generator files.

6.2.5 Numerical values for constant parameters in delay formulas

Recalling from Section 4.4.1, the delay constraint applies for the VBR applications in the downlink direction. Table 6.13 provides the values for the parameters used in the delay constraint formula.

Parameters	**Web browsing** High	**Web browsing** Low	**Mail**
$L^{max}_{f_s^{\text{DL\_WL}}}$	1460 byte	1460 byte	1460 byte
$L_{f_s^{\text{DL\_WL}}}$	800 byte	800 byte	800 byte
σ_s	1460 byte	1460 byte	1460 byte
Propagation time/link	20 ms	20 ms	20 ms
Transmission delay/node	10 ms	10 ms	10 ms
$D_s^{\text{DL-MAX}}$	2 s	5 s	10 s

TAB. 6.13: Value of parameters in delay formula

6.3 Network topology

In the network topology, the BSCs, external nodes and the wired links which connect the BSCs to external nodes are considered. We do not mention the wired links between the base stations and their corresponding BSCs, since those links will not be used unless the base station has been activated to serve at least one session. We assume that there is no activated base station in the beginning of the simulation. The used topology in this study is illustrated in Figure 6.2. This topology is called Network 02 in the whole chapter.

The specifications of this network topology are:

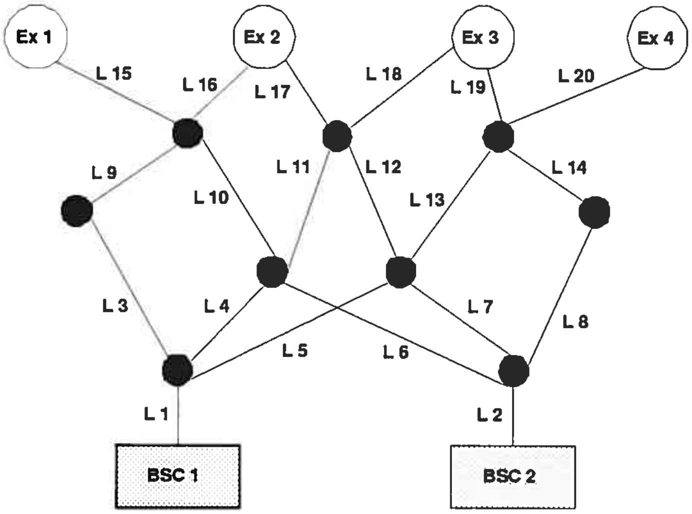


FIG. 6.2: Topology of the network

- Core network: 9 intermediates node and 18 wired links.
- Radio network: 2 BSCs and 2 wired links.
- Each path contains 4 links. When the serving base station of a session is chosen a link will be added to this path.

6.4 Definition of instances

In this study we have considered six different instances to compare the obtained objective values for the dimensioning of 3G networks in mono-routing and multi-routing. Each instance shows the number of potential base stations, number of sessions, network architecture, scenario and the applied application or applications in each particular simulation. For example, an instance is named "$02 : 30 - 200B.31$" means that the network architecture is of type Network 02, number of potential base stations is equal to 30, number of requested sessions is equal to 200 while all applications have been considered with scenario B. The instance "$02 : 30 - 300B.28$" means that this instance is again based on Network 02 topology while we have 30 potential base stations. The total number of sessions is 300 but we do not consider all applications. Here 28 means that we consider voice, web browsing, mail and videophone. In addition the number of sessions of each kind of application is calculated based on scenario B. For all instances, the soft handoff distance is equal to 150 meters. It means that the base stations which are located in 150 meters from the session can serve that in soft handoff and non soft handoff. Table 6.14 and Table 6.15 illustrate the instances used in this thesis. They contain different parameters of the instances, particularly the proposed RAB (data rate and FER) for each application and the proposed demanded Grade of Service for Gold and Silver priorities.

6.5 Optimization procedure: CPLEX

In order to solve the MIP problem of the model we use CPLEX-MIP, a software presented by ILOG [43] which solves linear and mixed-integer problems. CPLEX-MIP uses some parameters to solve the optimization problems. These are the parameters that can be personalized by each user for each individual problem and their functionality are explained in [41]. CPLEX-MIP uses a branch and cut algorithm where the problem is divided in different integer subproblems. CPLEX generates a branching tree. The root of this tree is the LP relaxation of the mixed-integer problem and each linear subproblem is a node of this tree.

The optimization procedure also uses cuts. Cuts are the added constraints to the model in order to eliminate some non-integer solutions. It happens when there are one or more fractional variables in the solution to the relaxation. The cuts also help to reduce the

		Instance02:30-200B.31			Instance02:30-200B.28			Instance02:30-200B.8		
Network Topology		Network 02			Network 02			Network 02		
Geographical surface		$500m \times 400m$			$500m \times 400m$			$500m \times 400m$		
Potential Base stations		30			30			30		
Number of Sessions		200			200			200		
Wired Links		50			50			50		
Paths		1 or 3			1 or 3			1 or 3		
Soft handoff Distance		150			150			150		
Grade of Service		Gold	0.1		Gold	-		Gold	0.1	
		Silver	0.3		Silver	0.3		Silver	0.3	
		RAB			RAB			RAB		
Applications	Dir.	Num.	r_t (kbps)(UL-DL)	FER(%)	Num.	r_t (kbps)(UL-DL)	FER(%)	Num.	r_t (kbps)(UL-DL)	FER(%)
Voice Gold	UL/DL	34	9.6	1	-	9.6	1	34	9.6	1
Voice Silver		85			119			85		
Videophone Gold	UL/DL	5	Low 76.8	0.5	-	Low 76.8	0.5	-	-	-
Videophone Silver		5	High 153.6		10	High 153.6		-		
Video streaming Gold	DL	6	Low 38.4	0.5	-	-	-	-	-	-
Video streaming Silver		11	High 153.6		-			-		
Web browsing Gold	UL/DL	9	Low (9.6-76.8)	5	-	Low (9.6-76.8)	5	9	Low (9.6-76.8)	5
Web browsing Silver		11	High (19.2-307.2)	5	20	High (19.2-307.2)	5	11	High (19.2-307.2)	5
Mail Gold	UL/DL	12	153.6	5	-	153.6	5	-	-	-
Mail Silver		22			34			-	-	-

TAB. 6.14: Instance I

		Instance02:30-300B.31			Instance02:30-300B.28			Instance02:30-400B.21		
Network Topology		Network 02			Network 02			Network 02		
Geographical surface		$500m \times 400m$			$500m \times 400m$			$500m \times 400m$		
Potential Base stations		30			30			30		
Number of Sessions		300			300			400		
Wired Links		50			50			50		
Paths		1 or 3			1 or 3			1 or 3		
Soft handoff Distance		150			150			150		
Grade of Service		Gold	0.1		Gold	-		Gold	-	
		Silver	0.3		Silver	0.3		Silver	0.3	
		RAB			RAB			RAB		
Applications	Dir.	Num.	r_t (kbps)(UL-DL)	FER(%)	Num.	r_t (kbps)(UL-DL)	FER(%)	Num.	r_t (kbps)(UL-DL)	FER(%)
Voice Gold	UL/DL	52	9.6	1	-	9.6	1	-	9.6	1
Voice Silver		127			179			239		
Videophone Gold	UL/DL	2	Low 76.8	0.5	-	Low 76.8	0.5	-	-	-
Videophone Silver		13	High 153.6		15	High 153.6		20		
Video streaming Gold	DL	7	Low 38.4	0.5	-	-	-	-	-	-
Video streaming Silver		18	High 153.6		-			-		
Web browsing Gold	UL/DL	17	Low (9.6-76.8)	5	-	Low (9.6-76.8)	5	-	Low (9.6-76.8)	5
Web browsing Silver		13	High (19.2-307.2)	5	30	High (19.2-307.2)	5	40	High (19.2-307.2)	5
Mail Gold	UL/DL	24	153.6	5	-	153.6	5	-	-	-
Mail Silver		27			51			-	-	-

TAB. 6.15: Instance II

number of branches along the tree to solve a problem.

If even after using cuts, the solution of the relaxation has still one or more fractional-valued integer variables, then CPLEX generates two new subproblems by branching on a fractional value with more restrictive bounds on the branching variable.

For instance in a minimization problem, the optimal solution of the continuous relaxation represents a lower bound for the optimal integer solution in which all the binary variables are equal to 0 or 1. Then the problem will be divided in two subproblems using a binary variable that is fixed either to 0 or 1. At this point CPLEX tries to solve the continuous relaxations of the subproblems. The result may be an integer solution, an infeasible solution or a fractional solution. In the last case, the process is repeated either by going deeper in the branching scheme or by jumping up in the tree. If the solution is integer the objective value defines an upper bound on the optimal value. This upper bound is compared with the incumbent one that is updated if necessary and the procedure continues on another branch. We can ignore a subtree if the obtained optimal solution on that subtree is greater than the incumbent value.

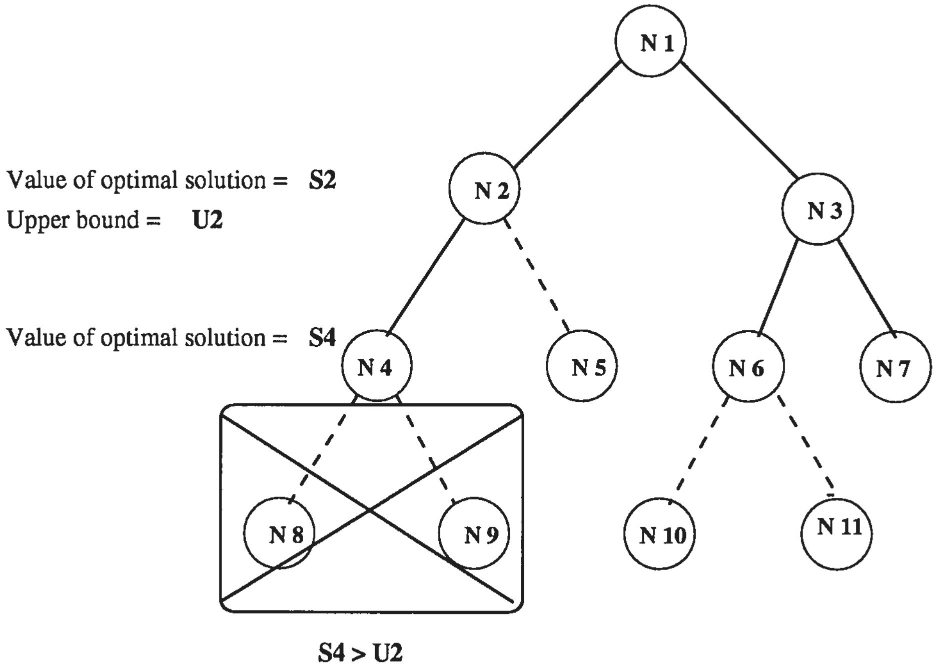

N 1
N 2
N 3
N 4
N 5
N 6
N 7
N 8
N 9
N 10
N 11
Value of optimal solution = S2
Upper bound = U2
Value of optimal solution = S4
S4 > U2

FIG. 6.3: Branching tree

Figure 6.3 illustrates an example. The value of optimal solution on node N4, (S4) is greater than the obtained upper bound on node N2, (U2) therefore we do not consider the children of node N4.

There are certain parameters by which we can control the branching scheme. The parameters taken in this study and their values are described in Table 6.16

6.6 Multi-routing profit over mono-routing

We expect to reduce significantly the value of the objective function when we consider several potential paths for each pair of source and destination nodes. This is what we call the profit of multi-routing over mono-routing. The proposed multi-routing strategy in this study with a set of potential paths for each session can help to accept real-time applications while satisfying their demanded QoS and reduce the usage of the resources in core and radio networks at the same time. The need of real-time applications to a high amount of bandwidth and their sensibility to transmission delay are the factors which make them more demanded comparing to the non-real-time applications. Therefore considering different paths instead of one path can help to reduce the used capacity even when the real-time applications are involved.

Here comes a small example which can illustrate the multi-routing profit over mono-routing. Assume that Figure 6.4 is a part of a core network where Ex_1 and Ex_2 are external nodes. During the first period BSC_2 tries to communicate with external node Ex_2 and during the second period BSC_1 tries to contact Ex_1. In mono-routing where the available path from BSC_1 to Ex_1 is the path "1" and the available path from BSC_2 to Ex_2 is the path "4", eight links are used and the capacity of all eight links are considered in the objective function.

On the other hand, in multi-routing we have a set of potential paths for each pair of source and destination nodes. In this example between BSC_1 and Ex_1 the set of potential paths is A={1,2} and between BSC_2 and Ex_2 this set will be equal to B={3,4}. Therefore in order to minimize the value of objective function the optimization procedure takes path 2 from the set A and path 3 from the set B. In this case the number

CPLEX parameter	Used Value	Description and default value
IloCplex::Cliques	2	-1 Do not generate clique cuts 0 Automatically determined 1 Generate clique cuts moderately 2 Generate clique cuts aggressively Default: 0
IloCplex::Covers	2	-1 Do not generate cover cuts 0 Automatically determined 1 Generate cover cuts moderately 2 Generate cover cuts aggressively Default: 0
IloCplex::GUBCovers	2	-1 Do not generate GUB cuts 0 Automatically determined 1 Generate GUB cuts moderately 2 Generate GUB cuts aggressively Default: 0
IloCplex::NodeSel	1	0 Depth-first search 1 Best-bound search 2 Best-estimate search 3 Alternative best-estimate search Default: 1
IloCplex::VarSel	0	-1 Branch on variable with minimum infeasibility 0 Branch variable automatically selected 1 Branch on variable with maximum infeasibility 2 Branch based on pseudo costs 3 Strong branching 4 Branch based on pseudo reduced costs Default: 0
IloCplex::BrDir	1	-1 Down branch selected first 0 Automatically determined 1 Up branch selected first Default: 0
IloCplex::EpAGap	different for each instance	Any positive number Default: 1e-06
IloCplex::EpGap	different for each instance	Any number between 0 and 1 Default: 1e-04

TAB. 6.16: Value of CPLEX parameters

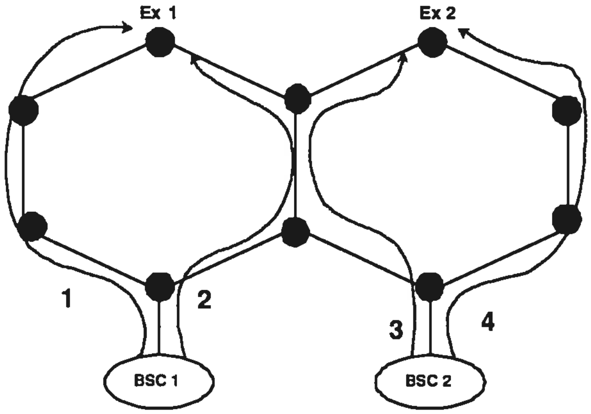


FIG. 6.4: An Example of core network

of used link reduces to seven.

This is the behavior that we expect from the optimization procedure, Figure 6.5.

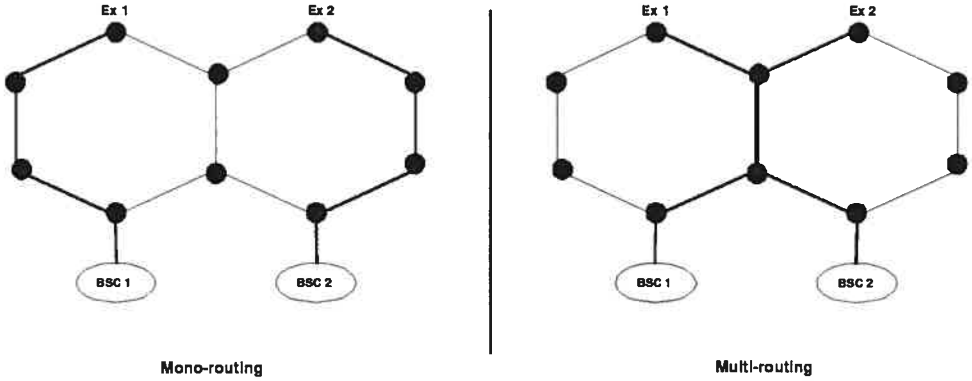


FIG. 6.5: Comparing mono-routing and multi-routing

6.7 LP value, best integer value and validation of the numerical results

The optimal value is the obtained value when the CPLEX parameters "EPGAP" and "EPAGAP" are equal to zero. If we consider a greater value than zero for these parameters, two values will be obtained. One as *best integer value* and the other as *LP value*. The optimal value lays between these two values.

$$LP\ value \leq Optimal\ value \leq Best\ integer\ value.$$

The smaller the gap parameters the closer the best integer value is to the optimal value. To make sure the profit of multi-routing over mono-routing, we consider the intervals of [LP value mono, best integer mono] and [LP value multi, best integer multi]. If there is no overlap between these intervals we can claim clearly a profit for the multi-routing strategy, Figure 6.6.

$$Best\ integer\ value\ multi \leq LP\ value\ mono.$$

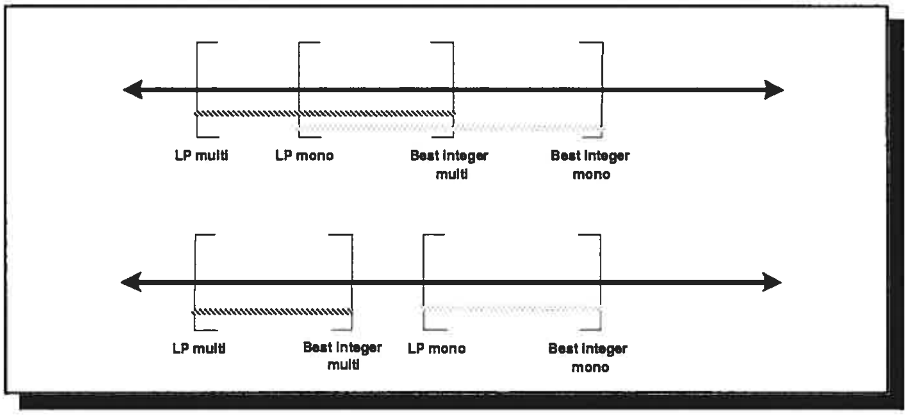


FIG. 6.6: Comparing the LP and best integer values in mono and multi-routing

6.8 Numerical results

6.8.1 Impact of multi-routing

In this thesis we have first studied the profit of multi-routing over mono-routing for all instances described in Table 6.14 and Table 6.15. The results presented in Table

	Instance	SHO	Number	Demanded	BS	Sessions	Capacity	Objective	LP
		Distance	of Paths	Gap				value	value
1	02:30-200B.31	150	1	5%	13	158	28476.8	34976.8	33264.7
2	02:30-200B.31	150	3	5%	12	158	27371.2	33371.2	31362.17
3	02:30-200B.28	150	1	1%	8	129	18924.8	22924.8	22708.1
4	02:30-200B.28	150	3	1%	8	129	18386.8	22396.8	22168.1
5	02:30-200B.8	150	1	0.05%	8	108	17116.8	21116.8	21103.8
6	02:30-200B.8	150	3	3%	8	108	16816	208116	19968.9
7	02:30-300B.31	150	1	10%	17	235	38620.8	47120.8	42543.8
8	02:30-300B.31	150	3	10%	16	235	36720	44720	40884
9	02:30-300B.28	150	1	10%	14	200	31200	38200	34525.7
10	02:30-300B.28	150	3	10%	13	194	30388	36888	33458
11	02:30-400B.21	150	1	10 %	15	210	31836.8	39336.8	35549
12	02:30-400B.21	150	3	10%	14	210	31334.4	38334.4	34122

TAB. 6.17: Comparing results when number of periods is 6

6.17 are obtained when the simulation time is divided in six periods. For all instances we can observe the profit of multi-routing over mono-routing looking at the objective values. Especially for two pairs of instances, number (3,4) and (5,6), in which we were able to run with relatively small gaps, there is no overlap between the best integer of multi-routing and the LP of mono-routing values, Section 6.7. If we go on solving the problem with a smaller gap (1e-06, 1e-04) then the LP and best integer values of mono-routing get closer and the LP value of multi-routing approaches the best integer value of multi-routing as well.

We can also note a decrease in the number of activated base stations in multi-routing compared to mono-routing. As it can be seen in pairs (1,2), (7,8), (9,10) and (11,12) the number of activated base stations decreases when we consider a set of potential paths instead of just the shortest path.

6.8.2 Impact of the number of periods

In order to evaluate the impact of the number of periods on the dimensioning we have increased the number of periods during the planning time. The new value for the number of periods is 10. Table 6.18 illustrates the obtained results for the three different instances. By comparing the results in Table 6.18 and Table 6.17 we notice a reduction in the value of the objective function.

	Instance	SHO Distance	Number of Paths	Demanded Gap	BS	Sessions	Capacity	Objective value	LP value
1	02:30-200B.31	150	1	5%	13	158	26946.4	33466.4	31650.14
2	02:30-200B.31	150	3	5%	12	158	26376	32376	30727.68
5	02:30-200B.8	150	1	0.005%	8	108	16870.4	20870.4	20546.34
6	02:30-200B.8	150	3	0.005%	8	109	16491.2	20491.2	20385
7	02:30-300B.31	150	1	10%	16	235	37686.4	43686.4	41580
8	02:30-300B.31	150	3	10%	16	235	35512	43512	39078.2

TAB. 6.18: Comparing results when number of periods is 10

We can justify this result by a small example. In Figure 6.7 we have considered a planning time first divided in 6 periods and then divided in 10 periods. The same sessions with same duration and beginning time are shown in both cases. Considering the first instance which has a duration equal to 5.8 seconds started on time 0:0, according to Section 4.2.1, when the total number of periods in the planning time is equal to 6, the resources are reserved for this session till the end of the second period. It means that this session reserves the resources for about 10 units of time. On the contrary, when the total number of periods during the planning time is 10, then the same session with the same duration and beginning time reserves the resources till the end of the second period among the 10 periods. It means that the same session but in the planning time divided in 10 periods reserves the resources for only 6 units of time. For the second session with a duration of 11 seconds, starting on 0:99 time, the resources are reserved all along the 3 first periods which corresponds to 15 units of time in the first case. While, in the second case where we have 10 periods the reservation take place during 12 units of time. The resource reservation for these sessions in 6 periods and 10 periods are illustrated in Figure 6.8.

In the case where we have divided the planning time in 6 periods there is an overlap between the first and third sessions in the use of resources. While the same two sessions

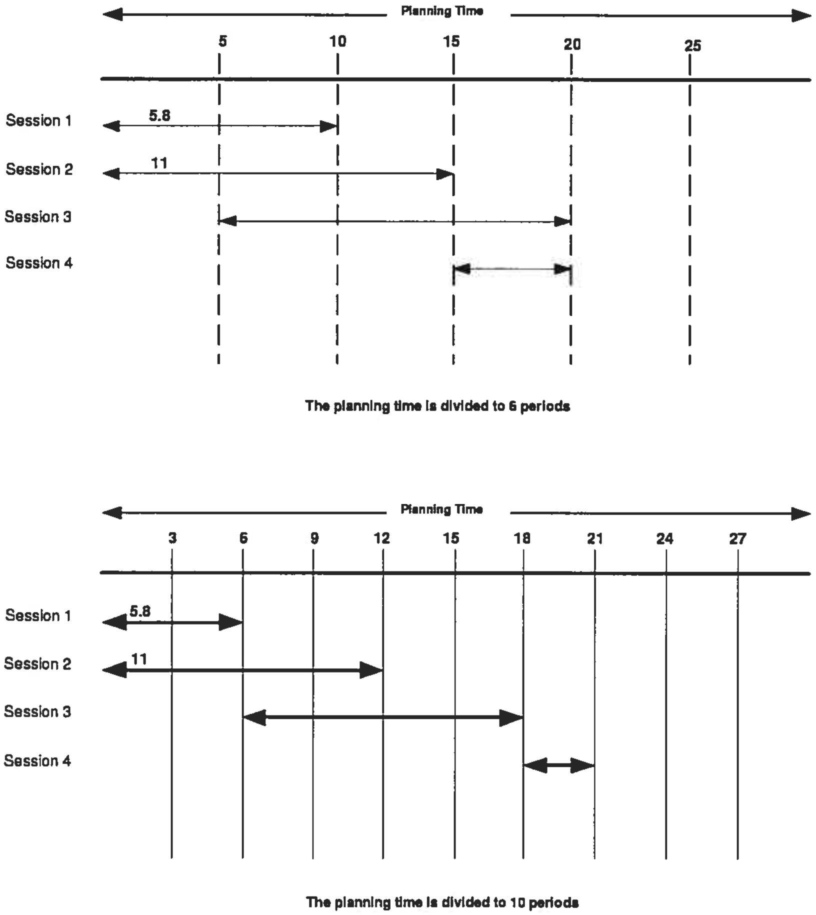


FIG. 6.7: Comparing sessions in 6 and 10 periods

do not have any overlap when we divide the planning time in 10 periods. It can be seen that, when we divide the planning time in 10 periods the resource reservation is less during the life of a session comparing to 6 periods. On the other hand, we are not allowed to increase the number of periods just to reduce the resource reservations as it has an impact on solving time of the problem, Section 6.1.

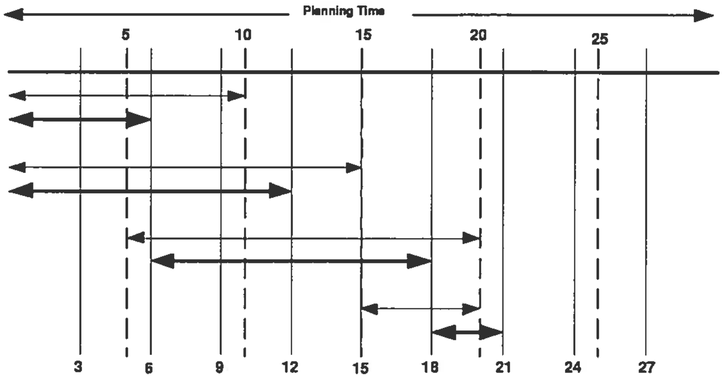


FIG. 6.8: Impact of increasing the number of periods during a planning time

6.8.3 Impact of the session length

In this section we study the impact of the session length on the dimensioning. We consider the same traffic files while the length of each session is divided by two, and a planning time which is divided in 10 periods. The shorter a session the less resources it reserves during the planning time. Table 6.19 illustrates that the values of the objective function decrease comparing to the values with the same number of periods in Table 6.18.

	Instance	SHO Distance	Number of Paths	Demanded Gap	BS	Sessions	Capacity	Objective Value	LP value
1	02:30-200B.31	150	1	5%	12	159	23779.2	29779.2	28266.5
2	02:30-200B.31	150	3	5%	11	158	231104.4	28610.4	27270.9
5	02:30-200B.8	150	1	0.005%	8	108	15536	19536	18959.05
6	02:30-200B.8	150	3	0.005%	7	108	15361.6	18861.6	18792.41
7	02:30-300B.31	150	1	10%	16	235	30764.8	38764.8	34971.9
8	02:30-300B.31	150	3	10%	15	235	29097.6	36597.6	33228.9

TAB. 6.19: Comparing results for the impact of sessions length

As it can be seen in pair instance (5,6) not only we have a multi-routing profit with respect to the capacity of the wired links, but also the multi-routing has an impact on the number of activated base stations. Note that for this pair of instances there is no overlap between the best integer value in multi-routing and LP value in mono-routing, Section 6.7. When we solve the problem with only one path which is also the shortest path, for serving instance 02:30-200B.8 the optimization procedure chooses 8 base stations among 30 potential base stations. While, for the same instance, this time with maximum three different paths, the optimization procedure chooses and activates 7 base stations.

6.8.4 Grade of Service

The proposed model respects the overall Grade of Service by accepting the demanded proportion of sessions in each type and priority of application during the planning time. This has been discussed in Section 5.4.6 where we have considered a blocking rate for different priorities (Gold and Silver) in each type of application. This blocking rate is global. It may happens that the blocking rate be violated during some periods while the global Grade of Service is obtained.

For instance, if we have been asked to accept at least 90% of the Gold voice sessions, it means that, at the end of the simulation time if we had 100 demanded Gold voice sessions we have to accept at least 90 of them.

For a traffic instance with 200 sessions of different types of application, when the planning time is divided in 10 periods, Table 6.20 illustrates the demanded and the obtained global Grade of Service.

The call admission control in this model is anticipative, therefore for accepting or rejecting a session it considers the traffic profile on the whole planning time and even

	Gold		Silver	
	Demanded GoS %	Obtained GoS %	Demanded GoS %	Obtained GoS %
Voice	95	97	80	90.66
Videophone	80	80	60	60
Video streaming	80	83.3	60	72.72
Web browsing	80	88	60	63.63
Mail	90	91.66	70	77

TAB. 6.20: Demanded and obtained GoS for all applications

the model may refuse a session while there is enough available resources to satisfy that session. This happens when the model has to accept another session with higher priority. Table 6.21 shows the number of demanded and accepted Gold and Silver voice sessions for the traffic instance with 200 sessions.

Based on this table, Figure 6.9 illustrates the functionality of the applied anticipative call admission control in details for each ten periods individually during the planning time for the voice application. The Grade of Service during a period is the rate of the accepted sessions which ask the service during that period.

The voice application has been chosen as we have new sessions asking for service in each period. Considering the Gold curve, there is a break in the curve during the third period while the model tries to compensate this by accepting more sessions over next periods in order to not violate the 95% demanded GoS. On the Silver curve, the break has happened during the fifth period and after that the model has tried to reach itself to the 100%. In this way the model will not violate the demanded GoS for Silver sessions as well.

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Demanded Voice Gold	7	3	2	5	5	4	4	1	-	3
Accepted Voice Gold	7	3	1	5	5	4	4	1	-	3
Demanded Voice Silver	8	5	7	9	8	12	11	4	10	11
Accepted Voice Silver	6	5	5	8	3	7	10	4	10	11

TAB. 6.21: Demanded and accepted number of voice sessions in each period

This behavior has been expected with the anticipative call admission control and the proposed objective function. The objective function aims at minimizing the resources, Hence it tries to accept as less sessions as possible, but it has no choice but satisfying the overall GoS.

The model has also considered the global GoS just to save the resources. If we force the system to accept 95% of Gold voices in each period then during the periods in which we have just 2 demanded sessions the system has no choice but accepting both of them. While it may sacrifice other Gold sessions, from other applications, because of lack of resources. Even in some cases we may not be able to solve the optimization problem.

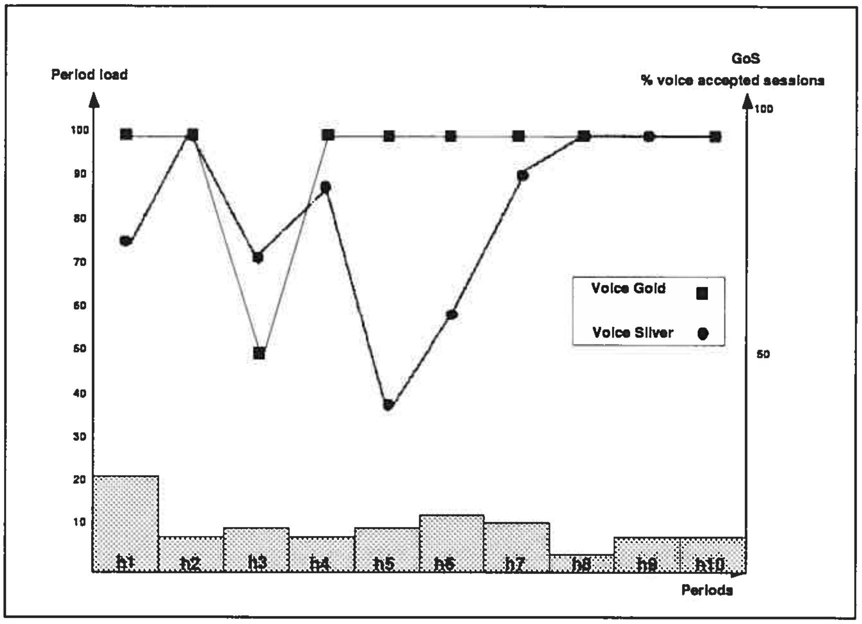


FIG. 6.9: Obtained GoS during each period for the voice application

On the contrary, we can consider a lower bound for each period as well. This will satisfy not only the overall Grade of service but also a lower bound on the Grade of Service during each individual period.

6.8.5 Conclusion

In this chapter we went through the steps taken since the mathematical model formulation which lead us to the obtained results. The various features of the program both in generating the traffic and generating the constraints have been explained in details. We also gave the values of the parameters used in the program and test the model with different instances under different conditions.
We sacrifice the number of constraints and variables in order to have a linear problem solved by CPLEX-MIP. As shown in Section 6.8 we were able to solve problem instances with up to 400 sessions and 30 potential base stations with a maximum of 3 different paths between each pair of source and destination on a surface of 400 m to 500 m.
In a planning time divided in 6 periods, we were able to show the profit of multi-routing

over mono-routing. The obtained dimensioning results illustrates that the model in multi-routing case reduces the used capacity of the wired links both in core and radio networks and the number of activated base stations in the radio network for the same number of accepted sessions.

Comparing the results of the same instances but with different number of periods during the planning time shows that the number of periods has an impact on the objective function. The more the number of periods the better the objective value. The impact of sessions length has been studied as well. The results show that not only the profit of multi-routing over mono-routing has been positive but also that the objective function decreases when the length of the sessions is divided by 2.

Chapter 7

Conclusion and Perspectives

The focus of this thesis was on the definition of an optimization model for the dimensioning of third generation networks while considering a multi versus a mono routing path policy. The study started from the model developed in C. Voisin M.Sc. thesis [1], made several improvements with respect to e.g., the traffic generator, the soft handoff features, the set of potential base stations for each session, QoS constraints, ... and generalized it for multi-routing strategies. The initial model included several mechanisms for satisfying the Quality of Service for each type of applications: an anticipative call admission control, a scheduling policy and a routing strategy based on a unique path for each source and destination pair of nodes. It was also integrating different kind of services and considered each session individually.

In this M.Sc. thesis, the whole study aimed to satisfy an objective function which concerns minimizing the dimensioning costs expressed by the number of activated base stations and the capacity of the wired links both in core and radio networks. In other words, it targets a minimum usage of radio and wired resources.

For this purpose the following steps have been followed:

➤ The future 3G networks and the multi-service traffic have been studied in order to model the core and radio networks, define the multimedia sessions with their demands and achieve an end-to-end quality of service.

- The dimensioning strategy considers the radio and wired part resources and chooses the best path to serve each accepted session based on an anticipative call admission control policy.
- The mathematical model corresponds to a linear mixed optimization problem solved using the CPLEX-MIP libraries of ILOG.
- Both the traffic generator and optimization tools, have been implemented using C++. The optimization tools include formatting the objective function, the set of constraints for solving the optimization model with CPLEX-MIP. The branching and exploration strategies have been adopted for solving each optimization problem on the CPLEX-MIP using the various parameters and options available in the CPLEX solver.
- Considering a network topology the dimensioning model has been used on different traffic instances to observe the impact of multi-routing over mono-routing on the use of the radio and wired resources.

Based on the obtained results we conclude:

- Comparing the results on mono-routing and multi-routing we can observe the profit of multi-routing in terms of used capacity on the wired links: this can be deduced based on the obtained results in Table 6.17 when the planning time is divided in 6 periods. Referring to the reasoning on Section 6.7 we can claim clearly a profit for the multi-routing strategy, see the third, fourth, fifth and sixth rows of Table 6.17.
- We can also observe the impact of multi-routing on the number of activated base stations: on fifth and sixth rows of Table 6.19 while the mono-routing LP value is bigger than multi-routing objective value we can observe the impact of multi-routing on the number of activated base stations.

- ➤ The results also show that the number of periods during the planning time influences the objective function and consequently the dimensioning of the multi-media network. This is based on the explanation in Section 6.8.2 and the obtained results in Table 6.17 and Table 6.18. For instance, comparing the first two rows in both tables we observe that the objective values when the planning time is divided in 10 periods are smaller than the objective values when the planning time is divided in 6 periods.

- ➤ The length of a session has an impact on the dimensioning, as it influences the reservation and ultimately use of the resources during the life time of the session. Considering two sessions with different lengths, the one with the shorter life time will be activated during a smaller number of periods. Consequently, comparing to the longer session its resource reservation will be less. This can also be seen in Tables 6.17, 6.18 and 6.19.

- ➤ The requested Grade of Service for each priority of application types has been respected all along the planning time. The objective function minimizes the use of the resources, therefore it tries to accept as less sessions as possible, while satisfying the Grade of Service.

- ➤ The results are meaningful even with the medium size network topology used in the computational experiments. If we change it to a network topology with more choice of paths for each pair of source and destination, the gain of multi-routing will appear even more positive.

As in any other model, some assumptions have been made:

- ➤ The traffic has been assumed homogeneous during the planning time on the geographical surface, while in the real life the traffic varies during a day. This also causes difficulties when the geographical surface is vast or if the planning time becomes longer.

- ➤ The temporal sequencing considered for each session leads us to an over-estimation of the session duration. In this model we reserve the resources for each session along more time than its real duration.

- The mobile stations are assumed to be fixed during the demands for a session: this prevents from the modeling of hand off for the mobile users.

- We have also considered that base stations are placed uniformly on the geographical space of simulation. Indeed this assumption of uniform positions of the base stations in the geographical space was due to the radio capacity formula that we used.

Future work:

- It could be interesting to consider a more complex network topology with more path choices for each pair of source and destination nodes. Meantime we should pay attention to the number of variables and constraints too, as it has an impact on the solution time of the mathematical program.

- The Call Admission Control used in this model is anticipative, which let us to evaluate the optimal utilization of the network. This policy helps more to satisfy the objective function than playing the real role of Call Admission Control. A Casual Call Admission Control with mono-routing, which takes into account just the current sessions in a period, has been developed in [2] where the results of the two models are compared. Comparing the results of these two call admission control policies in multi-routing and mono-routing would be interesting.

- Moves of the mobile station is another issue which can be studied in the dimensioning of the 3G networks. This will automatically lead us to dynamic multi-routing.

- A lower bound for the Grade of Service can be considered during each period for all type of applications, instead of a global GoS. Hence, we reduce the ups and downs of GoS during each period.

- A non homogeneous traffic is much realistic, therefore work can be done on radio capacity formulas for multi-services in order to adapt that to non homogeneous traffic.

Bibliography

[1] C. Voisin, *Définition d'un modèle d'optimisation pour le dimensionnement de réseaux troisième génération*, MSc Thesis, École Polytechnique de Montréal, Département de génie électrique, Decembre 2002.

[2] B. Jaumard, C. Meyer, R. Pooyania, Y. Solari and C. Voisin, *Causal and Anticipative Models for the Dimensioning of 3G Multi-service Networks*, Global Telecommunications Conference, 2003, GLOBECOM '03.IEEE, Vol. 6, pages 3417-3422, 1-5 December 2003.

[3] S. Clints and C. Daniel, *3G wireless networks*, McGraw-Hill, New York, 2002.

[4] A. Clapton, *Future mobile networks : 3G and beyond*, Institution of Electrical Engineers, London, 2001.

[5] H. Holma and A. Toskala, *WCDMA for UMTS*, Wiley, 2002.

[6] T. S. Rappaport, *Wireless Communications*, Prentice-Hall, 1996.

[7] Z. A. Uzmi, *Wireless Multiaccess Using CDMA*, Technical Report, Lahore University of Management Sciences, Computer Science and Engineering Department, Pakistan, September 2002.

[8] Benjamin Ip, *3G Wireless Network Architecture UMTS vs. CDMA2000*, ELEN 6951, Wireless and Mobile Networking II, Columbia University,2002, www.columbia.edu/itc/ee/e6951/2002spring/Projects/CVN.

[9] H. Kaaranen, A. Ahtiainen, L. Laitinen, S. Naghian and V. Niemi, *UMTS Networks Architecture, Mobility and Services*, Wiley, pages 15-99, 2001.

[10] K. Kasargod, M. Sheppard and M. Coscia, *Packet data in the Ericsson CDMA2000 radio access network*, Ericsson Review No. 3, pages 96-103, 2002.

[11] R. Ferrus and P. Diaz, *On hard/soft handoff and macro-diversity in CDMA mobile system*, PIMRC 99, Osaka, September 1999.

[12] 3GPP, 3rd Generation Partnership Project, Technical Group Services and System Aspects, QoS Concepts and Architecture (Release 1999), Tech. Rep. TS 23.107 v3.2.0, http://www.3gpp.org, 2000.

[13] P. P. White, *RSVP and Integrated Services in the Internet: A Tutorial* , Communications Magazine, IEEE, Vol. 35, Issue. 5, pages 100-106, May 1997.

[14] L. l. Peterson and B. S. Davie, *Computer Networks A Systems Approach*, Edition 3, Morgan Kaufmann Publishers, 2003.

[15] D. Levine, M. Naghshineh and I. Akildyz, *A Resource Estimation and Call Admission Algorithm for Wireless Multimedia Networks Using the Shadow Cluster Concept*, IEEE/ACM Transactions on Networking, Vol. 5, pages 1-12, February 1997.

[16] S. Keshav, *An Engineering Approach to Computer Networking*, ATM Networks, the Internet, and the Telephone Network, Addison-Wesley, 1999.

[17] S. Chen and K. Nahrstedt, *An Overview of Quality-of-Service Routing for the Next Generation High-Speed Networks: Problems and Solutions*, IEEE Network Magazine, Special Issue on Transmission and Distribution of Digital Video, Vol. 12, No. 6, pages 64-79, November-December 1998.

[18] S. Chen, *Routing Support for Providing Guaranteed End-to-End Quality of Service*, Ph.D. thesis, University of Illinois at Urbana Champaign, Department of computer scince, May 1999.

[19] Q. Ma and P. Steenkiste, *Quality of Service Routing for traffic with Performance Guarantees*, Proceedings of IFIP Fifth International Workshop on Quality of Service, pages 115-126, May 1997.

[20] R. Guerin and A. Orda, *QoS-based Routing in Networks with Inaccurate Information: Theory and Algorithms*, INFOCOM 97, Sixteenth Annual Joint Conference of the IEEE Computer and Communications Societies, Proceedings IEEE, Vol. 1, pages 75-83, 7-11 April 1997.

[21] Z. Wang and J. Crowcroft, *QoS Routing for Supporting Multi-media Applications*, Selected Areas in Communications, IEEE Journal, Vol. 14, Issue. 7, pages 1228-1234, September 1996.

[22] S. Chen and K. Nahrstedt, *On finding Multi-constrained paths*, Conference Record.1998 IEEE International Conference, Vol. 2, pages 874-8797, 11 June 1998.

[23] H. F. Salama, D. S. Reeves and Y. Viniotis, *A Distributed Algorithm for Delay Constrained Unicast Routing*, INFOCOM 97, Sixteenth Annual Joint Conference of the IEEE Computer and Communications Societies, Proceedings IEEE, Vol. 1, pages 84-91, 7-11 April 1997.

[24] I. Cidon, R. Rom and Y. Shivitt, *Multi-path Routing Combined with Resource Reservation*, INFOCOM 97. Sixteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings IEEE, Vol. 1, pages 92-100, 7-11 April 1997.

[25] S. Chen and K. Nahrstedt, *Distributed Quality of Service Routing in hight-Speed Networks Based on Selective Probing*, Local Computer Networks, LCN 98. Proceedings., 23rd Annual Conference, pages 80-89, 11-14 October 1998.

[26] ATM Forum, *Private Networks Network Interface (PNNI)*, V.1.0 Specifications, June 1996.

[27] D. Medhi and S. Guptan, *Network Dimensioning and performance of Multi-Service, Multi-Rate Loss Networks with Dynamic Routing*, IEEE/ACM Transactions on Networking, Vol. 5, pages 944-957, December 1997.

[28] D. Medhi and I. Sukiman, *Multi-service Dynamic QoS Routing Schemes with Call Admission Control: A Comparative Study*, Journal of Network and Systems Management, Vol. 8, No. 2, pages 157-190, June 2000.

[29] D. Medhi, *QoS Routing Computation with Path Caching: A Framework and Network Performance*, Communications Magazine, IEEE, Vol. 40, Issue. 12, pages 106-113, December 2002.

[30] Jun Jiang and Symeon Papavassiliou, *Providing End-to-End Quality of Service with Optimal Least Weight Routing in Next-Generation Multi-service High-Speed Networks*, Journal of Network and Systems Management, Vol. 10, No. 3, pages 281-308, September 2002.

[31] H. Abrahamsson, B. Ahlgren, J. Alonso, A. Andersson and P. Kreuger, *A Multipath Routing Algorithm for IP Networks Based on Flow Optimization*, International Workshop on Quality of future Internet Services (QofIS), Lecture Notes on Computer Science 2511, pages 135-145, Zurich, October 2002.

[32] H. D. Sherali, C.M. Pendyala and T.S. Rappaport, *Optimal location of transmitters for micro-cellular radio communication system design*, IEEE J.Select.Areas Commun., Vol. 14, No. 4, pages 662-673, 1996.

[33] K. Tutschku, *Demand-based Radio Network Planning of Cellular Communication Systems*, INFOCOM 98. Seventeenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE, Vol. 3, pages 1054-1061, 29 March-2 April 1998.

[34] C. Y. Lee and H. G. Kang, *Cell Planning with Capacity Expansion in Mobile Communications: Tabu Search Approach*, IEEE Transaction on Vehicular Technology, Vol. 49, Issue. 5, pages 1678-1691, September 2000.

[35] B. Jaumard, R. Malhame, O. Troeung, S. Esposito, C. Klam and C. Voisin, *Deliverable 2: Call Admission Control Policies and Dimensioning of Multimedia 3G Networks-Simulation Plan*, Technical Report, Ecole Polytechnique de Montreal, April 2002.

[36] S. Madan and M. Diez, *IS-2000 Forward Link Simulation. Performance Results*, Technical Report v6, Ericsson Documentation, 2000.

[37] S. Esposito, *Contrôle d'admission avec mesures pour une meilleure gestion des resources dans les réseaux de troisième génération*, MSc Thesis, École Polytechnique de Montréal, Département de génie électrique, April 2003.

[38] A. K. Parekh and R. G. Gallager, *A Generalized Processor Sharing Approach to Flow Control in Integrated Service Network: The Multiple Node Case*, IEEE/ACM Transaction on Networking, Vol. 2 No. 2, pages 137-150, April 1994.

[39] A. K. Parekh and R. G. Gallager, *A Generalized Processor Sharing Approach to Flow Control in Integrated Service Network: The Single-Node Case*, IEEE/ACM Transaction on Networking, Vol. 2 No. 2, pages 137-150, April 1994.

[40] B. Jaumard, C. Meyer and C. Voisin, *A Mathematical Model for the Dimensioning of 3G Networks*, In Preparation, Manuscript, 2003.

[41] ILOG, ILOG CPLEX 7.0 Reference Manual, 2000.

[42] B. Lind, *Application Traffic Model for UMTS Services* , Technical Report, PD4, Ericsson Documentation, 2000.

[43] ILOG. http://www.ILOG.com.

Appendix A

Leaky bucket

Leaky bucket is a popular scheme for traffic shaping in order to reduce the burstiness. The basic idea is to assign a bucket to each flow. Figure (A.1) shows a simple and clear idea of this mechanism.

Data is placed in a potentially infinite buffer on the left, while tokens are generated and placed in a bucket of size B at rate r. Means that there is no bound on the number of packets but the bucket contains at most B bits worth of tokens. To transmit a packet queued in the buffer, the regulator must first remove a number of tokens from the bucket corresponding to the packet size. This will make sure that a flow will never send faster than r worth of packets per second. It also guarantees that in an interval t the long term transmission rate is at most r.

The end to end delay bound, for a leaky bucket constrained session s computed as

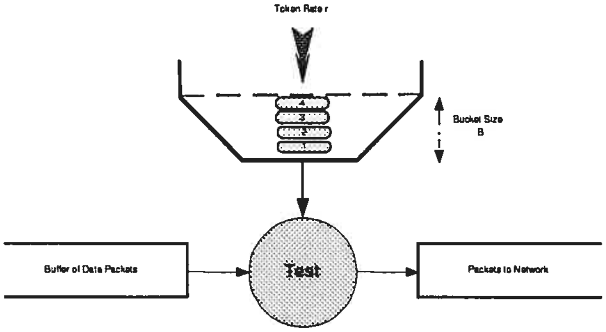


FIG. A.1: Leaky Bucket

follow:

$$D_s = \frac{B_s}{g_s} + \frac{(P_s - 1)l_s}{g_s} + \sum_{m=1}^{P_s} \frac{l_{max}}{w_m} \tag{A.1}$$

Where g_s is session s guaranteed rate and should not be less than the r_s which is the average token bucket rate, l_s is the maximum packet length of session s, l_{max} is the maximum packet length in the network, P_s the number of hops in the path and w_m is the total bandwidth for the m-th hop.

Appendix B

Selective Repeat Protocol

Selective Report error recovering protocol is a procedure to transmit the damaged or missed frames on radio links and provides reliability. Frames are numbered in this protocol. This specification helps to keep the frames in order. The protocol uses NAKs (Negative Acknowledgment) for damaged and lost frames. Selective Repeat has two windows, one for sending and one for receiving parts. The sending window defines which frames are outstanding. However the sent frames would be kept in receiving windows. The frames are buffered until their predecessors arrive. That is why the size of memory in receiving part should be big enough to keep the frames.

When an out of order frame arrives, the protocol detects a damaged or missed frame and send a NAK for the missed frame. The NAK notifies the sender of the loss. At this point the sender will stop sending the frames in order, and will resend the frame specified by NAK.

We consider a transmission time for each frame even the NAKs. Therefore if we ignore the queuing time of a frame, the delay will be equal to the sum of NAK and retransmission times. The lost of the the retransmitted frames happen in radio network. In these cases the protocol will wait a certain amount of time to send a second NAK. In this study we have calculated the average time of transmission of a frame when maximum three retransmissions are needed.

B.1 Average time of transmission

- T_{frame} : Time of transmission of a frame
- $D^{transmission-radio}$: The average time of transmission of a frame
- D^{k}_{frame} : Delay time of a frame when k transmission is done
- T_{NAC}: Time between two consecutive NAK frames
- P : The probability that a frame is in error with the selected RAB. We call this FER

The probability that k transmission attempts are needed to successfully transmit a frame is:

$$P^k = P^{k-1} \times (1 - P) \tag{B.1}$$

The average transmission time when maximum three retransmission is needed will be:

$$D^{transmission-radio} = \sum_{k=1}^{4} P^k \times D^k_{frame}. \tag{B.2}$$

Therefore we will have:

$$D^{transmission-radio} = P^1 \times D^1_{frame} + P^2 \times D^2_{frame} + P^3 \times D^3_{frame} + P^4 \times D^4_{frame}, \tag{B.3}$$

$$\begin{aligned} D^{transmission-radio} = {} & (1 - FER) \times T_{frame} \\ & + FER \times (1 - FER) \times 3 \times T_{frame} \\ & + FER^2 \times (1 - FER) \times 3 \times T_{frame} \times T_{NAC} \\ & + FER^3 \times (1 - FER) \times 3 \times T_{frame} \times 2 \times T_{NAC}, \end{aligned} \tag{B.4}$$

If we consider $T_{NAC} = 2 \times T_{frame} + 60$ with $T_{frame} = 100$ ms we will have $T_{NAC} = 260$

FER	$D^{radio-transmission}_{T_{FER}}$
0.5%	101.01 ms
1%	102 ms
2%	104.11 ms
5%	110.68 ms
10%	122.78 ms

ms. Based on this information the average transmission time with different Frame Error Rate is calculated in the following. The $D_{r_{\mathrm{FER}}}^{radio-transmission}$ is used in the delay formula in Chapter 4.