TECHNICAL UNIVERSITY OF CRETE

ELECTRONIC & COMPUTER ENGINEERING
DEPARTMENT

TELECOMMUNICATIONS DIVISION

Diploma Thesis

“TCP Performance over UMTS Network”

Smaragdakis Georgios

Submitted to the Department of Electronic and Computer Engineering in partial fulfillment of the requirements for the degree of Diploma of Engineering in Electronic and Computer Engineering at the Technical University of Crete

Supervisory Committee:

Professor Paterakis Michael (Supervisor)
Professor Digalakis Vasilios
Professor Sidiropoulos Nikolaos

στην οικογένειά μου,
Θεμιστοκλή, Άννα και Εμμανουήλ.

ABSTRACT

Universal Mobile Telecommunications System (UMTS) is the forthcoming global mobile network for packet data. This network uses the Wideband Code Division Multiple Access (WCDMA) air Interface. Contrary to other Radio Networks like GPRS, most of UMTS applications will be end-to-end applications and as a result the Transmission Control Protocol (TCP) will be used.

In this thesis, at first, we describe this network (both UMTS Radio Access Network and UMTS Core Network) and use it as the platform to study performance, behavior and energy consumption of TCP over it. Several TCP versions are discussed including TCP Tahoe, TCP Reno, TCP New Reno, TCP Vegas and TCP Sack. We examine these TCP versions using well-known application layer protocols running over TCP, like FTP, Telnet and HTTP. Our simulation results show that there is not an optimal TCP version for all applications, however, one outperforms the others in terms of energy consumption which is critical in UMTS. As base of our simulations we use Network Simulator.

ACKNOWLEDGEMENT

These five years was a fabulous journey. I just can not imagine myself not taken part in it. On the other hand this half decade of my life was not easy and without disappointments. This is why I would like to thank some people who inspired me and encouraged me to continue my “trip to Ithaki”.

At first I would like to thank my family and especially my brother who supported me to overcome any problem and was permissive enough to tolerate my acidulated character. Moreover I would like to thank my good friend Maria K. for her love, support and encouragement the last four years. My friends Dimitris K.,George K., Akis K., Eleni F. It is my honor to be your friend.

I would also like to thank the faculty of my department for their effort. Especially my supervisor, Professor Paterakis M. Without his support this thesis would have never end.

I am planning to continue my journey to a distant, new “Ithaki”…

… cu later … I promise.

PREFACE

Chapter 1 is an introductory Chapter and is addressed to those who know little about Mobile Networks.

In Chapter 2 we describe the evolutions from GSM to GPRS and then to UMTS. We describe which parts of UMTS are entirely new and which of them are inherited from elder technologies. Moreover we explain why in UMTS added services follow an end-to-end approach.

Chapter 3 is one of the largest chapters in the thesis as it describes the lossy part of UMTS, the WCDMA air interface. At first we describe WCDMA and then some major error causes: Path Loss, Shadowing and Wideband Effect. We then describe RAKE receiver which is a technology used in the UMTS in order for the receiver to synchronize all the received signals. Moreover we describe the UMTS Radio Resource Management and Handoffs. Furthermore we explain how important power control is for UMTS and describe the major Power Control Techniques. Last but not least, we analyze how cell capacity is estimated.

In Chapter 4 we describe the Core Network Architecture in order to emphasize the end-to-end functionality of UMTS and we then describe the application classes of UMTS.

In Chapter 5, the Transport Control Protocol (TCP) is analyzed in some detail. We describe the TCP Reliable Data Transfer and Retransmission Mechanism as well as TCP flow control, TCP Round Trip and timeout, TCP connection Management and TCP congestion control. We then describe the five versions of TCP examined in this thesis. We explain TCP latency modeling and describe UDP. At the end of the Chapter, we discuss why TCP can be a headache when we use it over wireless links.

In Chapter 6 we present the system parameters used in our simulation. We then we explain how the error model is introduced in our simulation for each user, using two state and three state Markov models. Moreover we present improvements for TCP over the UMTS network and describe the Network Simulator, a discrete simulator which is the base of our simulator. Last but not least we give traffic specifications of our simulations.

In Chapter 7 we present and discuss our results. We examine performance and energy consumption of the various versions of TCP for the differnet applications. Moreover we examine the efficiency of each application related to each TCP version.

In the last chapter, Chapter 8, we present the conclusions of our work and propose ideas for future work and extensions.

CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

TABLE OF GRAPHS

CHAPTER 1: Mobile Networks

1.0 Introduction

Mobile networking is one of the most prominent areas of networking today. Subscribers of mobile services grow at a rate of 10 millions new users per month and as a result new mobile networks have to be developed in order to satisfy the additive needs of the entire new wireless environment of communications and business. In order to understand this mobile evolution we have to come through the terminology and basic principles of a mobile network, which is the global standard for - the most popular - terrestrial mobile communications.

1.1 Architecture

Figure 1.1 [3] shows a sample mobile network with three Base Transceiver Stations (BTS), one Base Station Controller (BSC) and one Mobile Switching Center (MSC).This figure also shows four Mobile Stations (MSs). In a typical network there are thousands of BTSs. The BTS is also called Radio Base Station (RBS).

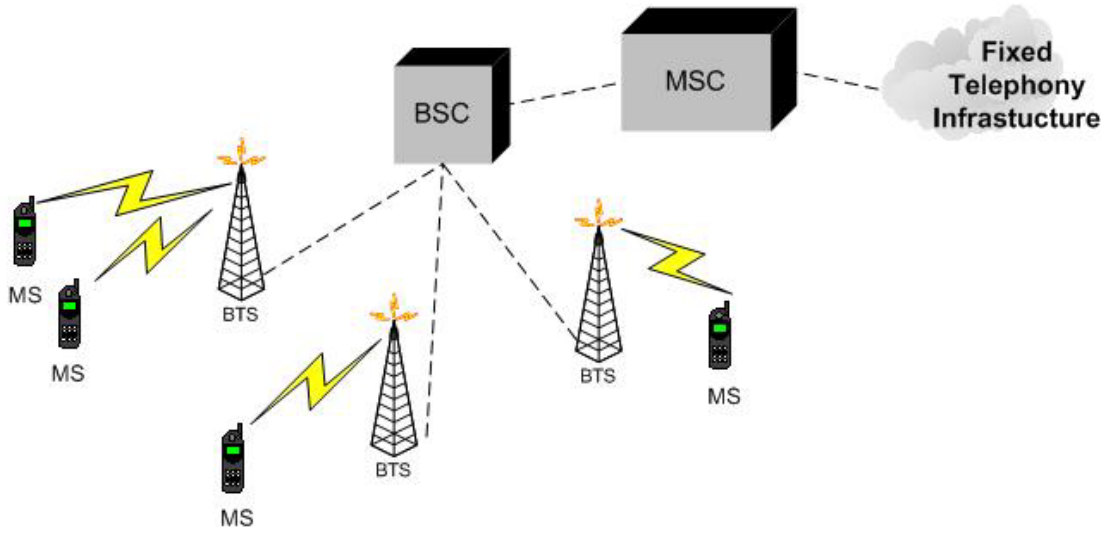


Figure 1. 1 : Sample Mobile System Architecture

1.1.1 The Base Station Subsystem

Although the architecture varies a bit between different systems, there is always an antenna that receives signals from the handsets and transports then to the mobile systems. The antennas are usually located in various high level places in order to obtain the best possible coverage. Connected to each antenna is usually a base station that processes the call setups and routes the calls to the network. In Figure 1.2 the base station is depicted as an antenna tower, although there is special equipment nearby the tower which is responsible for the communications functionality.

A cell is the basic geographical unit of a cellular system and is defined as the area of radio coverage that one base station antenna system provides. Each cell is assigned a unique number called a Cell Global Identity (CGI). The coverage area of a mobile system consists of a huge number of these cells, hence the words cellular system and cellular phones.

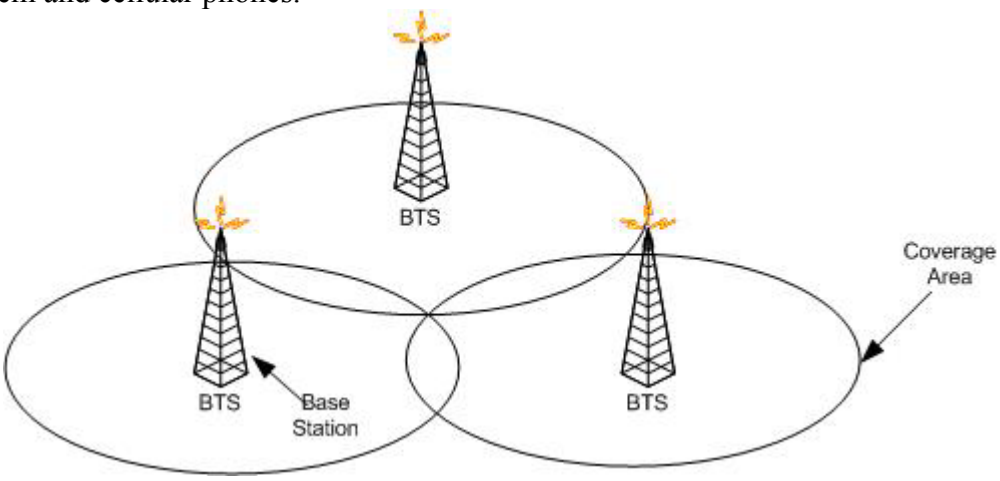


Figure 1. 2 : Cells and Antennas

One cell sometimes sends information in all directions from the base station, and sometimes there are three sectors surrounding the antenna. The first configuration is common in rural areas, where it is crucial to obtain as high coverage as possible. The latter configuration, on the other hand, is especially suited for high traffic areas, and the cells can be directed in clever ways in order to cope with the high traffic. In these cases, one cell is usually aimed directly at that spot so that it does not deal with any other traffic. So, a base station has an antenna that enables an air interface connection with the MS. When setting up a call, there are commonly some resources (transceiver, power and so on) allocated to the user in question.

A number of base stations are then connected to a controller BSC in the case of the Global System for Mobile Communication (GSM) and to a Radio Network Controller (RNC) in the case of Wideband Code Division Multiple Access (WCDMA). Much of the intelligence of mobile system exists here. The BSC/RNC manages all advanced radio related functions, handover (moving from one cell to another), radio channel assignments, Quality of Service (QoS), and the collection of cell configuration data. Advanced load balancing and admission control functionality also exists in the BSC/RNC. The controllers and the base stations together are referred to as the base station subsystem.

1.1.2 The Core Network

The core network has traditionally been equipped with switches and subscriber handling functionalities. These features include subscriber handling, authentication, security, and system maintenance. As more and more advanced services are introduced, the core network becomes more and more of a data network in which circuit-switched and packet-switched services share the same network. Nowadays, this migration is becoming more obvious. The main task of the traditional core network is to route traffic that enters a mobile network from other networks to the right base station and to route calls from an MS within the system to the right destination, as shown in Figure 1.3.

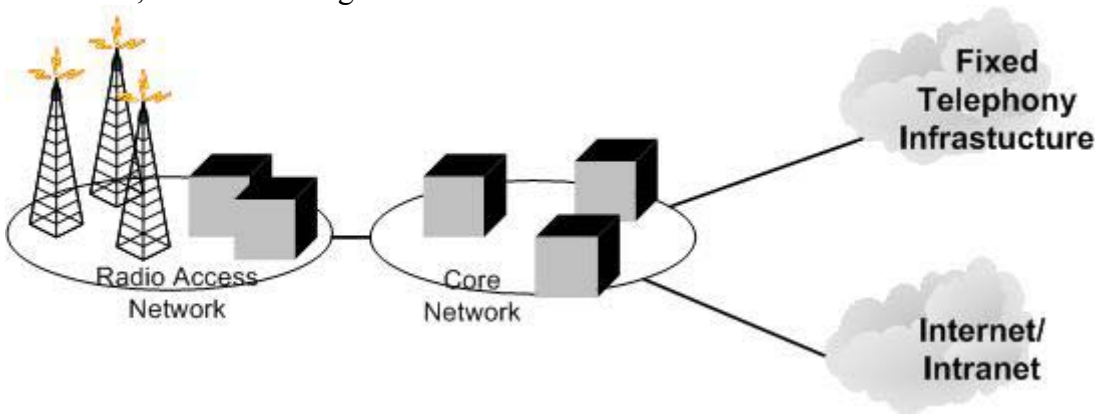


Figure 1. 3 : The Core Network transports voice and data to and from the Radio Access Network

The destination network for data services might be another mobile network, a land line phone network, or the Internet. The advent of advanced data services changes this situation, however, and creates a need for items such as Short Message Service (SMS) centers, Application gateways and so on.

1.1.3 Other Networks

After our call is routed from the MS via the base station, the BSC and the core network, it now finds the right destination network. The core network switch determines whether the call should be sent to a land line phone network, to another mobile phone network, or to a different destination. If the destination network is a mobile system, this route is repeated in reverse order. At the base station, the MS is paged with a signal that tells it that someone wants to reach it.

Therefore phones do not talk directly with each other; rather, they communicate via networks. The base stations do not send the calls directly to each other; instead, they communicate via a network that most of the time is buried in the ground. (Note: by a “call” we refer to any demand for mobile communication services)

1.2 Concepts and Terminology

The terminology we use consist of the widely accepted telecommunication jargon and of some concepts that are specific to the emerging mobile Internet industry. The latter often lack clear definitions, and different sources use them in various ways. This section aims to remove those ambiguities and to provide a set of concepts that we can consistently use throughout this thesis.

1.2.1 Separating Users from Each Other

In a mobile system, different users need to use different channels in order to avoid traffic collisions. The three most common ways to achieve this goal are frequency division, time division, and code division. Frequency Division Multiple Access (FDMA) gives each user a different frequency. The first analog systems commonly used FDMA. Time Division Multiple Access (TDMA) separates users in time by assigning different time slots for each channel. Each channel is called a time slot because it allocates a certain time interval during each radio frame. In GSM, there are eight time slots in each frame, giving each user the opportunity to send every eight time slots. Code Division Multiple Access (CDMA) is used by the majority of the Third Generation (3G) systems, as well as cdmaOne. In CDMA, different users are separated by different codes. CDMA requires very good power control algorithms, or else only the loudest users could be heard [3].

1.2.2 Separating Sending and Receiving Traffic

In telecommunications, the world uplink and downlink are often used to describe outgoing and incoming traffic for the handset, respectively. Figure 1.4 illustrates these two concepts. Furthermore, the choice of duplex method determines how uplink and downlink traffic for one user is separated. Time Division Duplex (TDD) separates the uplink and downlink channels in time. This is used by Bluetooth for instance [3]. Frequency Division Duplex (FDD) allocates different frequencies for the uplink and downlink channels. WCDMA/FDD is an example of how different frequencies are used for sending and receiving.

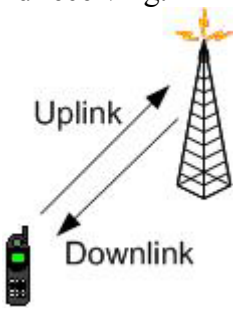


Figure 1. 4 : Uplink and Downlink

CHAPTER 2: Basics of UMTS Network

2.0 Introduction

Beginning in 1998 major partners (including the European Telecommunications Standard Institute (ETSI)) started discussions aiming at coopering for the production of standards for a third generation mobile system with a core network based on evolutions of GSM an access network based on all the radio access technologies (both frequency- and time-division duplex mode) supported by different partners. This project was called the Third Generation Partnership Project (3GPP). Almost one year later the American National Standard Institute (ANSI) decided to establish 3GPP2 a 3G partnership project for evolved ANSI/Telecommunications Industry Association (TIA)/Electronics Industry Association (EIA)-41 networks. There is also a strategic group called International Mobile Telecommunications (IMT-2000) within the International Telecommunication Union (ITU), which focuses on defining interfaces between 3G networks evolved from GSM on one hand and ANSI-41 on the other hand, in order to enable seamless roaming between 3GPP and 3GPP2 networks. Thanks to this universal (worldwide) roaming characteristic, 3GPP started referring to 3G mobile systems as Universal Mobile Telecommunication System (UMTS). In 3GPP, the UMTS specification work was divided into two phases. For the first phase of UMTS, Release 1999 (R99), standardization work was finished around the end of 1999 and the beginning of 2000. This phase was more or less a logical evolution from the second generation system architecture. The second phase, called Release 2000 (R00), is a complete revolution introducing many new concepts and features [2].

UMTS will offer a common air interface that will cover all fields of application and have the flexibility to integrate worldwide the different mobile communications systems. UMTS will be the first system to offer mobile users roaming during an existing connection, with handover between networks with different applications and different operators. With the system it will be possible to transmit voice and data over one connection and subscribers will be assigned a personal telephone number that will allow them to be reached anytime, anywhere in the world [4].

2.1 Evolution from GSM to UMTS

Evolution is one of the most common terms used in the context of UMTS. It is the technical evolution that is how and what equipment and in which order they are brought to the existing network if any [11]. Moreover evolution as a high-level

context covers not only the technical evolution of network elements but also expansions to network architecture and services. During this process it must be clear that a network is as strong as its weakest element and due to open interfaces defined in the specifications many networks are combinations having equipment provided by many vendors. On the other hand service evolution is not such a straightforward issue. It is mainly based on end user's demands. In the following sections we will try to briefly describe, the major steps and technology evolution from the simple GSM system to the complicated UMTS.

2.1.1 The GSM

Global System for Mobile Communications (GSM) was the first fully digital mobile telecommunications system, known as 2nd Generation system (2G). The main idea behind GSM specification was to define several open interfaces which determine the standardized components of the GSM system. Furthermore GSM specification in principle provided the means to distribute intelligence throughout the network. From the GSM point of view, this decentralized intelligence is implemented by dividing the whole network into four separate subsystem; Network Subsystem (NSS), Base Station Subsystem (BSS), Network Management Subsystem (NMS) and Mobile Station (MM), as shown in the following figure.

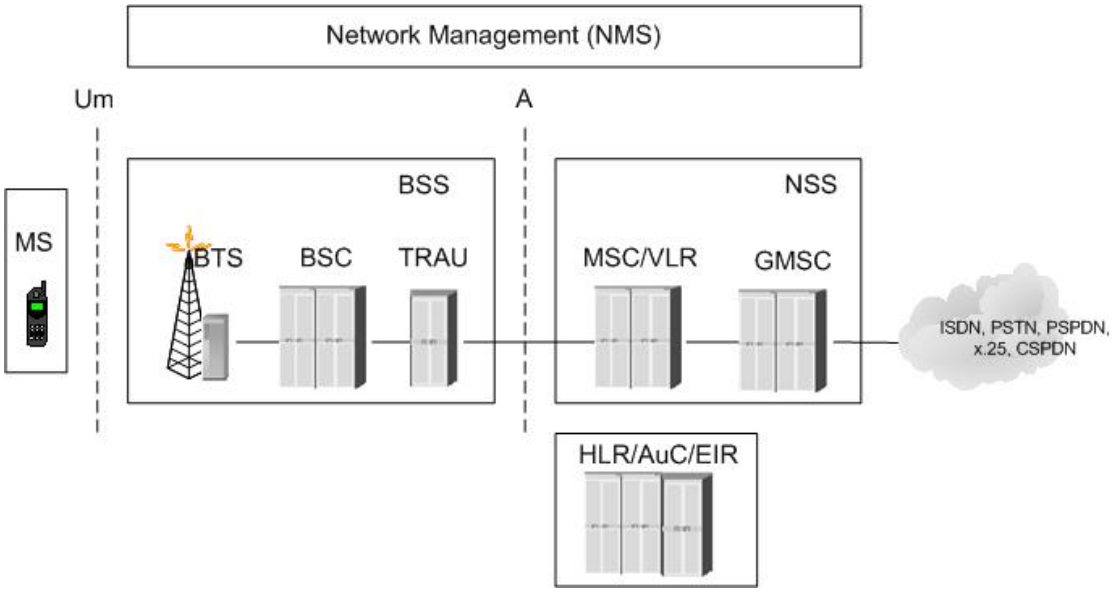


Figure 2. 1: Basic GSM network and its subsystem

The actual network needed for cell establishment is composed of the NSS, the BSS and the MS. The BSS is a network part taking care of all call control functions. Every call is always connected by through the NSS. The NMS is the operation and maintenance related part of the network. It is also needed for the whole network control. The network operator observes and maintains the quality and services of the network through the NMS. The open interfaces in this concept are related between the MS and BSS (Um interface) and between the BSS and NSS (A interface). The interface between the NMS and the NSS/BSS was expected to be open, but its

specifications were ready in time and this is why every manufacturer implements NMS interfaces with their own proprietary methods.

The MS is a combination of terminal equipment and a subscriber's service identity module. The terminal equipment as such is called Mobile Equipment (ME) and the subscriber's data is stored in a separate module called the Service Identity Module (SIM). Hence, ME+SIM = MS.

The Base Station Controller (BSC) is the central network element of the BSS and it controls the radio network. This means that the following functions are BSC's main responsibility areas: maintaining radio connections towards the MS and terrestrial connections towards the NSS. The Base Transceiver Station (BTS) is a network element maintaining the air interface (Um interface). It takes care of air interface signaling, ciphering and speech processing. In this context speech processing means all the methods BTS performs in order to guarantee an error-free connection between the MS and the BTS. The Transcoding and Rate Adaptation Unit (TRAU) is a BSS element taking care of speech Transcoding i.e. is capable of converting speech from one digital coding format to another and vice versa.

The Mobile Services Switching Center (MSC) is the main element of the NSS from the cell control point of view. MSC is responsible for call control, BSS control functions, internetworking functions charging, statistics and interface signaling towards BSS and interfacing with the external networks (PSTN/ISDN/packet data networks). Functionally the MSC is split into two parts, though these parts could be in the same hardware. The serving MSC/VLR is the element maintaining the BSS connections, mobility management and internetworking. The Gateway MSC (GMSC) is the element participating in mobility management, communication management and connections to the other networks. The Home Location Register (HLR) is the place where all the subscriber information is stored permanently. The HLR also provides a known, fixed location for the subscriber-specific routing information. The main functions of the HLR are subscriber data and service handling, statistics and mobility management. The Visitor Location Management (VLR) provides a local store for all the variables and functions needed to handle calls to and from mobile subscribers in the area related to the VLR. Subscriber related information remains in the VLR as long as the mobile subscriber visits the area. The main functions of the VLR are subscriber data and service handling and mobility management. The Authentication Center (AuC) and Equipment Identity Register (EIR) are NSS network elements taking care of security-related issues. The AuC maintains subscriber identity-related security information together with the VLR. The EIR maintains mobile equipment identity (hardware) related security information together with VLR.

2.1.2 Value Added Services Platform

The very natural step to develop the basic GSM was to add service nodes and services centers on top of the existing network infrastructure. The GSM specifications define some interfaces for this purpose, but the internal implementation of the service centers and nodes are not the subject of those specifications. The common name of the service provided by the centers and nodes is Value Added Services (VAS) Platforms and this term describes quite well the main point of adding this equipment to the network as shown in the figure 2.2. The minimum VAS platform contains typically two pieces of equipment; Short Message Service Centre (SMSC) and Voice

Mail System (VMS). Technically speaking the VAS platform equipment is relatively simple and is meant to provide a certain type of service. They use standard interfaces towards the GSM network and may or may not have external interface towards other networks.

To massive these services to the end-users, a more individual type of services is required. To make this possible, the Intelligent Network (IN) concept was integrated together with the GSM network. Technically this means major changes in the switching network elements in order to add the IN functionality and in addition, the IN platform itself is a relatively complex entity. IN enables service evolution to take big steps towards individuality and also with IN the operator is able to perform more secure business, for example, pre-paid subscriptions are mostly implemented with the IN technology.

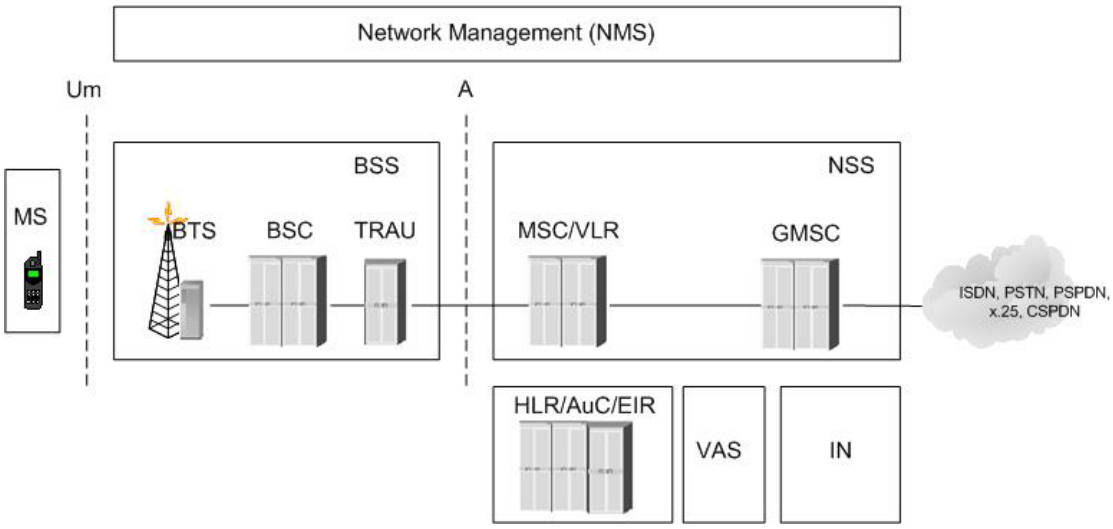


Figure 2. 2: Value Added Service Platform

In the beginning, GSM subscribers have used 9.6Kbps circuit switched symmetric pipes for data transfer. Due to the Internet and electronic messaging the pressure for mobile data transfer has increased a lot and this development was maybe underestimated at the time when the GSM system was specified. To ease this situation a couple of enhancements have been introduced. Firstly, the channel coding is optimized. By doing this the effective bit rate has increased from 9.6 to 14kbps. Secondly to put more data through the air interface, several traffic channels can be used instead of one. This arrangement is called High Speed Circuit Switched Data (HSCSD). In an optimal environment a HSCSD user may reach a data transfer rate of 40-50kbps. The higher data rates are usually required only in the downlink, but unfortunately this solution is symmetric. Hardware and Software changes must be incorporated in the terminal (MS), BTS, BSC, MSC/VLR and HLR/AuC/EIR.

2.1.3 2.5 Generation

The circuit switched symmetric Um interface is not the best possible access media for data connections. When also taking into account that the majority of data

traffic is packet switched in nature, something more had to be done in order to "upgrade" the GSM network to make it more suitable for more effective data transfer. The way this was done is referred to as the General Packet Radio Service (GPRS). GPRS requires two additional mobility network specific service nodes: Serving GPRS Support Node (SGSN) and Gateway GPRS Support Node (GGSN). By using these nodes the MS is able to form a packet switched connection through the GSM network to an external packet data network (e.g. the Internet). This approach is also called Second and Half Generation (2.5G).

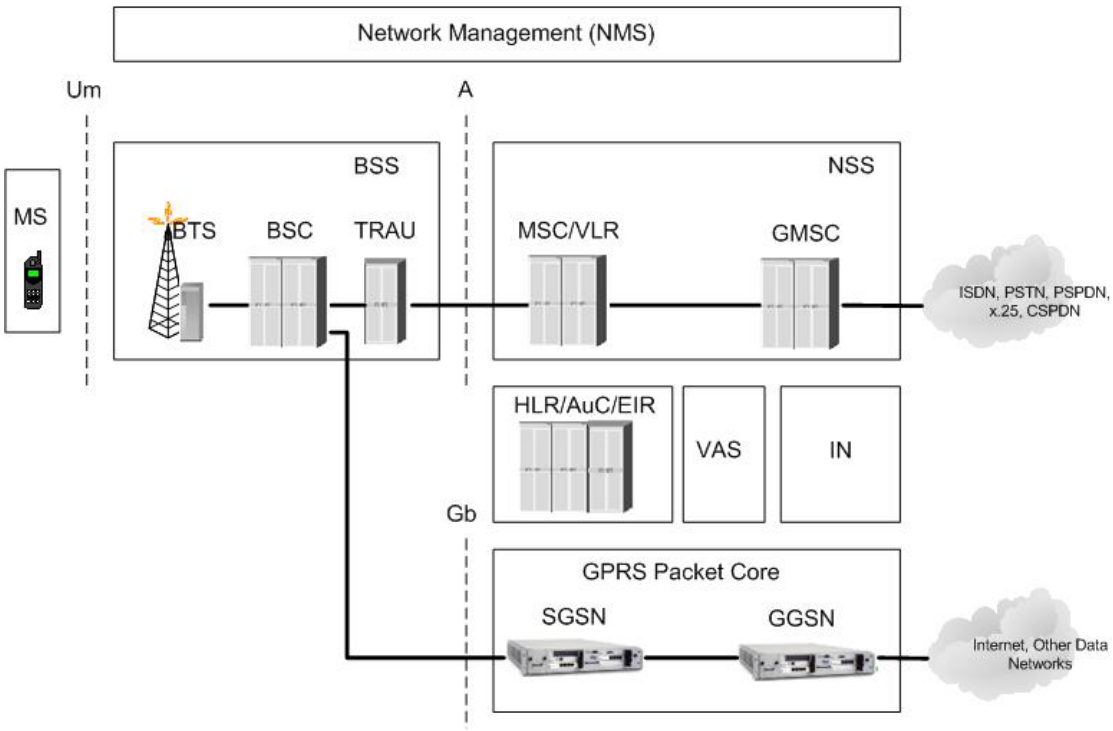


Figure 2. 3: GPRS Service

GPRS has the capability to use asymmetric connections when required and thus the network resources are better utilized. GPRS is a step bringing IP mobility and the Internet closer to the cellular subscriber, but it is not a complete IP mobility solution. From the service point of view, GPRS start a development path where more and more services traditional supported by GSM circuit switching are converted to be offered used over GPRS because those services were originally more suitable for packet switched connections. One example of this is Wireless Application Protocol (WAP), the potential of which is to be discovered when offered over GPRS.

When packet switched connections are used, the QoS is a very essential issue. In principle the GPRS supports the QoS concept but in practice it does not. The reason here is that GPRS traffic is always second priority traffic in the GSM network: it uses otherwise unused resources in the Um interface. Because the amount of unused resources is not exactly known in advance, no one can continuously guarantee a certain bandwidth for GPRS and thus QoS cannot be guaranteed either.

By applying a completely new air interface modulation technique, Octagonal Phase Shift Keying (8-PSK), where one air interface symbol carries a combination of three information bits, the bit rate in the air interface can be remarkably increased. When this is combined with very sophisticated channel coding technique(s), one is able to achieve a data rate of 48kbps compared to conventional GSM which carries 9.6kbps per channel and one information bit in one symbol in the air interface. These

technical enhancements are referred to as the Enhanced Data Rates for Global/GSM Evolution (EDGE).

Development of the EDGE concept is divided into two phases. EDGE Phase1 and EDGE Phase2. EDGE Phase1 is also known as E-GPRS (Enhanced GPRS). Also the BSS is renamed as e-RAN (EDGE Radio Access Network). EDGE Phase1 defines channel coding and modulation methods that enable up to 385kbps data rates for packet switched traffic under certain conditions. The assumption here is that one GPRS terminal gets 8 air interface slots that is 8x48kbps. In addition, the EDGE terminal must be located close to the BTS in order to use a higher channel coding rate. EDGE Phase2 contains guidelines on how this speed is achieved for circuit switched services. EDGE Phase2 is also commercially known as E-HSCSD.

From the network evolution point of view, EDGE in general has its pros and cons. A good point is the data rate(s) achieved; there are almost equal to UMTS urban coverage requirements. The disadvantage with EDGE is that the data rates offered are not necessarily available throughout the cell. If EDGE is to be offered with complete coverage, the number of cells will remarkably increase. In other words, EDGE may be an expensive solution in some cases. The future of EDGE is still to be seen since it has to compete with the true 3G solutions.

2.2 3rd Generation

Third Generation (3G) introduces the new radio access method, Wideband CDMA. WCDMA and its variants are global , hence all 3G networks should be able to accept access by any 3G network subscriber. In addition to globality, WCDMA has been extensively studied in laboratory premises and it has been realized that it has better spectral efficiency than TDMA, in certain cases, and that it is more suitable for packet transfer than TDMA based radio access, WCDMA and radio access equipment as such are not compatible with GSM equipment, and this is why adding the WCDMA to the network one must add new elements: Radio Network Controller (RNC) and Base Station (BS) [11].

On the other hand, one of the key requirements for UMTS is GSM/UMTS interoperability. One example of interoperability is inter-system handover, where the radio access changes from GSM to WCDMA and vice versa during the transaction. The interoperability is taken care of by two arrangements. First, the GSM air interface is modified so that it is able to broadcast system information about the WCDMA radio network in the downlink direction. Naturally the WCDMA radio access network is able to broadcast system information about surrounding the GSM network in the downlink direction too. Second, to minimize the implementation costs, the 3GPP specifications introduce possibilities to arrange inter working functionality with which the evolved 2G MSC/VLR becomes able to handle the wideband radio access, UTRAN.

So far the abilities provided by the IN have been enough from the service point of view. The concept of IN is directly adopted from the PSTN/ISDN networks and thus it has some deficiencies as far as the mobile user is concerned. The major problem with standard IN is that the IN as such is not able to transfer service information between networks. In other words, if a subscriber uses IN based services

they work well but only within his or her home network. This situation can be handled by using “evolved IN” call Customized Applications for Mobile network Enhanced Logic (CAMEL). CAMEL is able to transfer service information between networks. Later on, the role of CAMEL will increase a lot in 3G implementation , actually almost every transaction performed through the 3G network will experience CAMEL involvement at least to some extent.

Transmission connections within the WCDMA radio access network are implemented by using Asynchronous Transfer Mode (ATM) on top of a physical transmission media (in 3GPP R99 Implementation). A pre-standardization project discussed a lot whether to use ATM in the network or not. The final conclusion was to use ATM because of the following two reasons:

1. ATM cell size and its payload are relatively small. The advantage here is that the need of information buffering decreases. When a lot of buffering is needed, expected delays will easily increase and also the buffering equipments of the corresponding will increase. One should bear in mind that buffering and thus associated packet delays have a negative impact on the QoS requirements of real-time traffic.
2. The other alternative, IP, and its IP version 4 (IPv4) was also considered but IPv4 has some serious drawbacks, being limited in its addressing space and missing QoS. On the other hand, ATM and its bit rate classes match very well with QoS requirements. This leads to the conclusion that where ATM and IP are combined (for packet traffic), IP is used on top of ATM. This solution combines the good points of both protocols: IP qualifies the connections with the other networks and ATM takes care of the connection quality and also routing. Due to the IPv4 drawbacks, a compromise has been made. Certain elements of the network use fixed IPv4 types of address but the real end-user traffic uses dynamically allocated IPv6 addresses, which are valid within the 3G network. To adapt the 3G network to the other networks in this case, the 3G IP backbone network must contain an IPv4 to IPv6 address conversion facility, because the external networks may not necessarily support IPv6.

The core network nodes are evolved technically, too. The Circuit Switched (CS) domain elements are able to handle both 2G and 3G subscribers. This requires changes in MSC/VLR and HLR/AuC/EIR. For example, security mechanisms during the connection set-up are different in 2G and 3G and now these CS domain elements must be able to handle both of them. The Packet Switched (PS) domain is actually an evolved GPRS system. Though the names of the elements here are the same as those in 2G, their functionality is different. The most remarkable changes concern the SGSN, whose functionality is very different from that in 2G. The SGSN is mainly responsible for Mobility Management activities for a packet connection. In 3G the Mobility Management entity is divided between the RNC and SGSN. This means that every cell change the subscriber undergoes in UTRAN is not necessary visible to the PS domain, but RNC handles this situation.

There will be three types of 3G mobile terminals (A,B and C) which are more complex than ME and from now on we call them User Equipment (UE) : circuit-switched, packet-switched (IP) and those that support both modes. Both circuit- and packet-switched modes are supported at the radio interface. The circuit-switched mode is used for the traditional circuit-switched terminals and makes optimal use of the radio resources for voice-services. Circuit-switched voice is optimized in terms of

both bandwidth (small frame protocol overhead) and quality. The packet-switched mode is more suitable for multimedia applications, but is less efficient in terms of bandwidth consumption due to the IP header over the radio.

2.2.1 Third Generation System Release 1999 (3GPP R99)

The 3G network implementation according to 3GPP R99 offers the same services with those of GSMPhase2+ (GPRS/EDGE). That is, all the same supplementary services are available, teleservices and bearer services have different implementation but this is not visible to the subscriber; a speech call is still a speech call, no matter whether it is done through a traffic channel (GSM) or by using 3G bandwidth . In addition to GSM, the 3G network in this phase may offer some other services not available in GSM, for example, video call could be one of those. In this phase the majority of services are moved/transferred/converted to PS domain whether reasonable and applicable. WAP is one of those candidates, because the nature of the information transferred in WAP is packet switched. The PS domain is taken into effective use and one service branch containing a variety of different services will be location based services utilizing the subscriber location mechanisms built into 3G network.

The development steps after 3GPP R99 are somehow unclear today, but some major trends are visible. The main trends in the following development steps are separation of connection, its control and services and at the same time, the conversion of the network to be completely IP based. From the service evolution point of view, these development steps recognize that multimedia services should be provided by the 3G network itself. Multimedia means a service where at least two media components are combined for example voice and image.

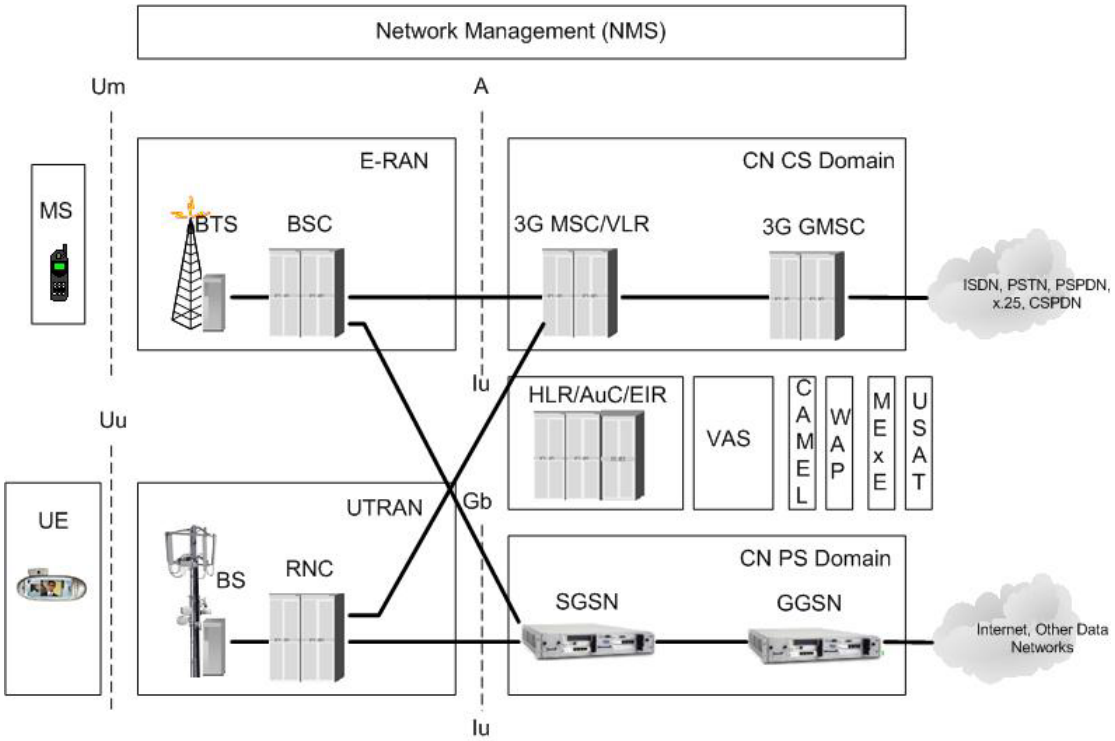


Figure 2. 4: 3GPP R99 network scenario

2.2.2 Third Generation System Release 4 (3GPP R4)

The above trends are big issues as such and this is why they are implemented in phases, the first phase being Third Generation System Release 4 (3GPP R4). The 3GPP R4 implementation introduces separation of connection, its control and services for the CN CS domain.

In the CN CS domain actual user data flow goes through Media Gateways (MGW), which are elements maintaining the connection and performing switching functions when required. The whole process is controlled by a separate element evolved from MSC/VLR called MSC server. One MSC server can handle numerous MGWs and thus CN CS domain is scalable; when one wishes to add control capacity, an MSC server is added. When one desire to add switching capacity, MGW's are added.

When this kind of network has been set up, the pace of technology development and specifications set the next limit. The more momentum the IPv6 gains, the more of the 3G network connections that can be converted to IPv6 too. This decreases the need of IPv4 to IPv6 network conversions. In this phase the relationship between circuit and packet switched traffic will remarkably change. The majority of the traffic is packet switched and also some traditionally CS services, like speech, will become at least partially packet switched (Voice over IP – VoIP) although these applications can be characterized as killer applications. To implement these applications a new CN subsystem called IP Multimedia Subsystem (IMS) is added here since it will offer unified methods to perform IP based multimedia services. Naturally the BSS part of the network could be implemented using IP but the time schedule of this change is unclear. The role of CAMEL will change too. Because many of the services using CAMEL are converted from the circuit switched side to the packet switched side of the network, CAMEL will now have connections also to the PS domain elements. In addition to this CAMEL will be connecting elements between the service platforms and the network.

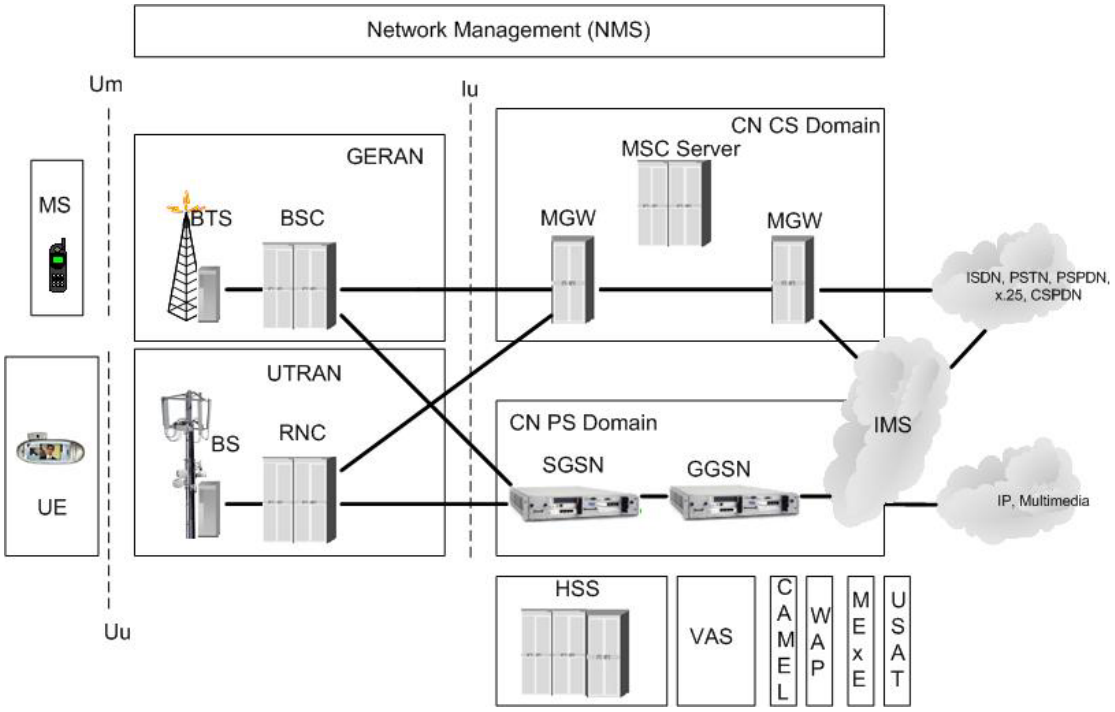


Figure 2. 5: 3GPP R4 implementation scenario

2.2.3 Third Generation System Release 5 (3GPP R5)

In Third Generation System Release 5 (3GPP R5) also known as All IP 3G, the evolution continues further and all traffic coming from the UTRAN is assumed to be IP based. If we think of a voice call from UE to PSTN as an example, this will be transported through UTRAN as packets and from the GGSN the VoIP is routed to the PSTN via IMS, which provides the required conversion functions.

From the UE point of view the network always looks the same in the development phases illustrated in figures 2.4 to 2.6. Inside the network almost everything changes. The major change will occur in transport technology, which in 3GPP R99 implementation is ATM. The 3GPP R4 and R5 implementation scenarios aim to swap ATM for IP. Because the system must be backward compatible, the operator always has a choice whether to use ATM or IP as the transport technology or whether the solution will contain both technologies. As explained earlier, the strength of ATM is its natural support for QoS. However, as time goes by IP will contain QoS mechanisms implemented over various kinds of sub networks, not only ATM.

In this phase the services and use of the network are more important than technology itself and due to this the used access technology may become less important. The main selection criterion for the used radio access technology is to offer enough bandwidth for the services used. The future vision here is that the 3G CN will have interfaces for several radio access technologies, for instance, GSM, EDGE, CDMA2000, WCDMA and Wireless Local Area Networks (WLAN). Naturally this imposes many requirements on the terminal manufactures and terminals capable of handling different kinds of access technologies will be introduced according to the market needs.

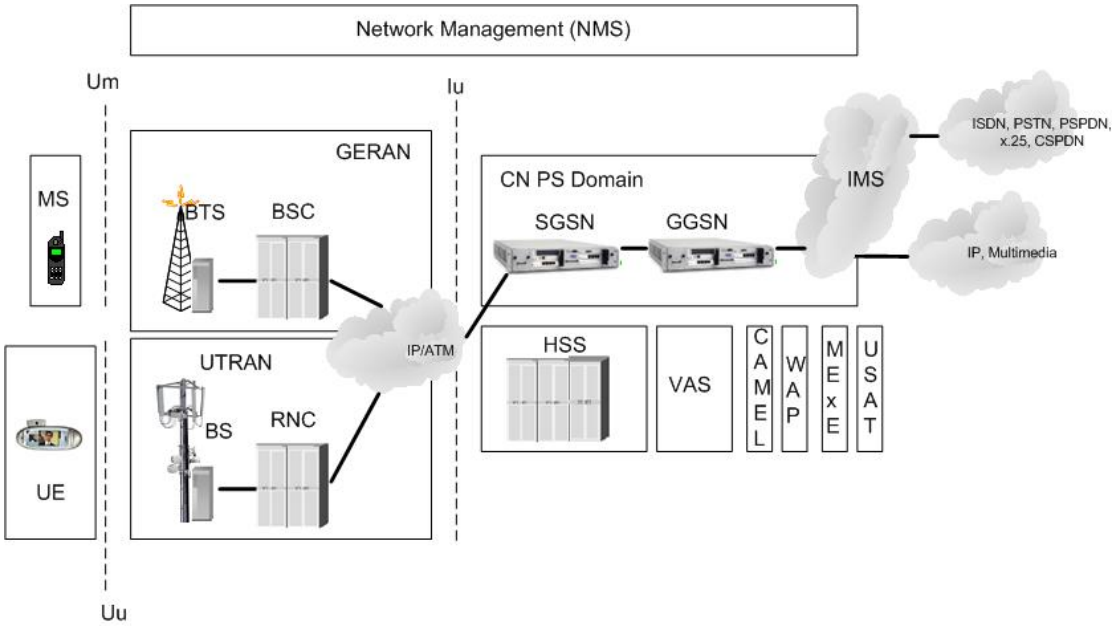


Figure 2. 6: Vision of 3GPP R5 (All IP)

CHAPTER 3: The UMTS Radio Access Network

3.0 Introduction

The main goal of the UMTS Radio Access Network (UTRAN) is to create and maintain Radio Access Bearers (RAB) for communication between UE and CN. With RAB the CN elements are given an illusion about a fixed communication path to the UE, thus releasing them from the need to take care of radio aspects. Referring to the network architecture models presented before, UTRAN realizes certain parts of QoS architecture independently as shown in the next figure.

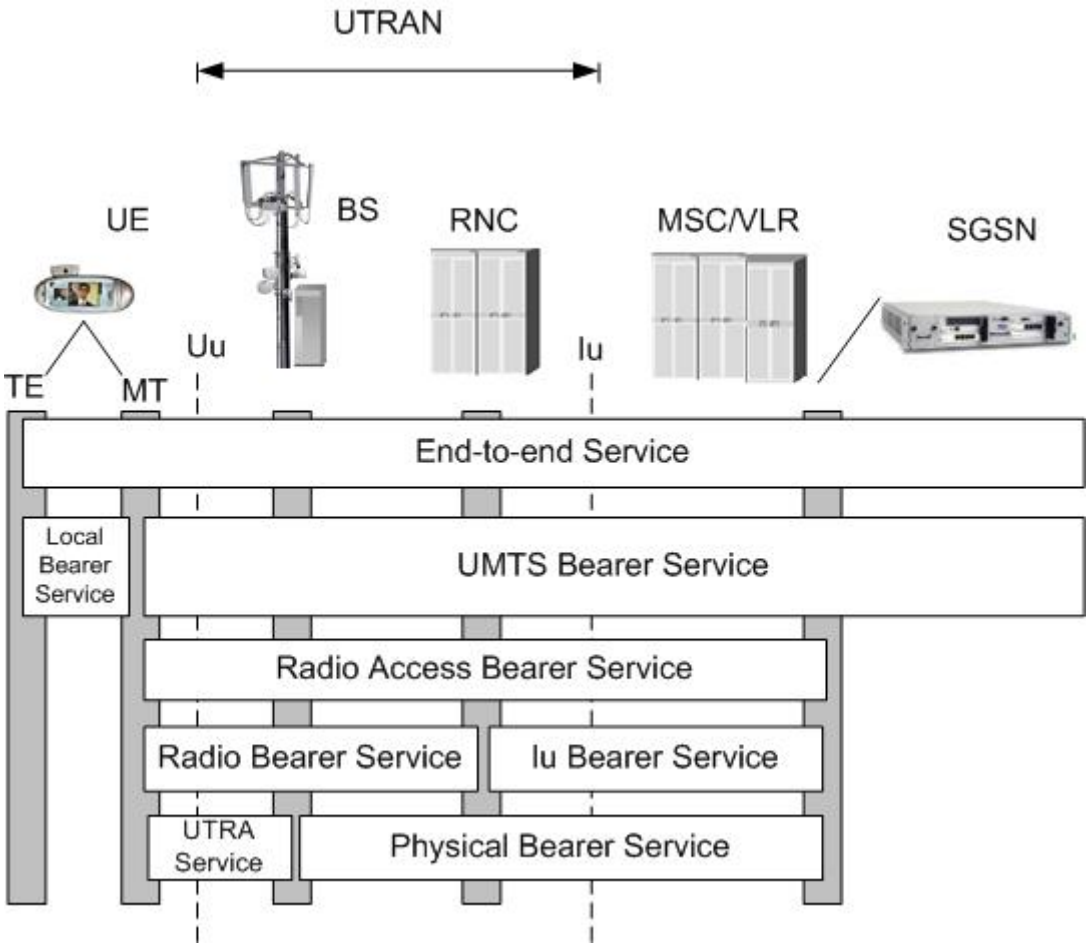


Figure 3. 1: Bearer/QoS architecture in UTRAN

UTRAN is located between two open interfaces, Uu and Iu. From the bearer architecture point of view, the main task of UTRAN is to provide bearer service over these interfaces; in this respect the UTRAN controls Uu interface and in Iu interface the bearer service provision is in co-operation with the CN.

The RAB fulfils the QoS requirements set by the CN. The handling of the end-to-end QoS requirement in the CN and in the UE is the responsibility of communication management. Those requirements are then mapped onto the RAB, which is "visible" for the Mobile Termination (MT) and CN. As said earlier, the main task of UTRAN is to create and maintain a RAB so that the end-to-end QoS requirements are fulfilled.

One of the main ideas of this layered structure is to encapsulate the physical radio access, later it can be modified or replaced without changing the whole system. In addition it is a known fact that the radio path is very complex and continuously changing transmission medium. This bearer architecture gives remarkable role to the RNC, since the RNC and the CN map the end-to-end QoS requirements over the Iu interface and the RNC takes care of satisfying the QoS requirements over the radio path. These two bearers exist in the system because the Iu bearer is more sable in nature; the Radio Bearer (RB) experiences more changes during the connection. For example, one UE may have three continuously changing RBs maintained between itself and the RNC, still the RNC has only one Iu bearer for this connection. This kind of situation occurs in context with the soft handover which is described later.

The physical basis of the end-to-end service in Uu interface is Universal Terrestrial Radio Access (UTRA) service. As was explained earlier, UTRA service is implemented with Wideband Code Division Multiple Access (WCDMA) radio technology and in the beginning of UMTS the used variant was WCDMA-FDD.

3.1 Basics of UMTS Radio Communications

The most popular CDMA system is WCDMA. The standard WCDMA for the UMTS is the FDD:DS-WCDMA. In Direct Sequence CDMA (DS-CDMA) scheme, the data signal is scrambled by the user specific pseudo noise (PN) code at the transmit side, for example at the mobile or base station, to achieve the spreading of the signal with the desirable chip rate and process gain. At the receiver (mobile or base station), the original signal is extracted by exactly the same spreading code sequence. As a result, every signal is assumed to spread over the entire bandwidth of the radio connection. Interference may therefore be generated from all directions in contrast to narrow-band cellular systems.

This type of WCDMA is the most prominent mode, it uses spectrum at 1920-1980 MHz for the uplink and 2110-2170 MHz for the downlink. This band is mostly described as 2GHz band. In WCDMA users are separated with code sequences. All users share the same frequency (5 MHz carrier) and time. All data both control data and user traffic are transmitted simultaneously. Uplink and downlink are frequency separated (2x5 MHz). Different users can use different services. One user can use different services at the same time eg. videophone and voice. The bit-rate of a connection can change during the connection's lifetime. In WCDMA the QPSK code

modulation is used. In the following figure we illustrate the frequency distribution in a WCDMA system.

The following table summarizes the main characteristics of FDD: DS-CDMA scheme described above (some explanation about these parameters are given in next paragraphs):

Parameter	Specifications
Multiple Access	FDD: DS-CDMA
Duplex Scheme	FDD
Chip Rate (Mc/s)	3.84
Frame length (ms)	10
Channel Coding	Convolution coding (R=1/2, 1/3, 1/4 , k=9); turbo code of R=1/2,1/3,1/4 and k=4
Interleaving	Inter/intraframe
Data Modulation	FDD: DL:QPSK, UL dual channel QPSK
Spreading modulation	FDD: UL:BPSK, DL:QPSK
Power Control	Closed Loop (inner loop and outer loop), open loop Step size: 1-3 dB (UL); power cycle: 1500/s
Diversity	RAKE in both BS and MS; antenna diversity; transmit diversity
Inter-BS synchronisation	FDD: no accurate synchronization needed
Detection	MS&BS: pilot symbol based coherent detection in UL, CPICH channel estimation in DL
Multiuser detection	Supported (not at the first phases)
Service Multiplexing	Variable mixed services per connection is supported
Multirate concept	Is supported by utilizing variable spreading factor and multicode
Handover	Intra-frequency soft and softer handovers are supported, inter-system and inter-frequency handovers are supported

Table 3. 1 : WCDMA –FDD: main technical characteristics

The spreading factor determines how large a code you will use when spreading the data. The chip rate is system specific and indicates the rate of bits that are sent over the air. For WCDMA, the chip rate is 3.84Mcps or 3840 bps as a bit over the air is commonly called a chip. Table 3.2 illustrates Spreading Factors for Different Bit Rates [3]:

USER BIT RATE (Kbps)	SPREADING FACTOR	CHIP RATE (Mcps)
30	128	3.84
60	64	3.84
120	32	3.84
240	16	3.84
480	8	3.84
960	4	3.84
1920	2	3.84

Table 3. 2 : Spreading Factors for Different Bit Rates

3.1.1 Path Loss and Shadowing

Both Path Loss and Shadowing are long-term propagation effects. Path Loss relates the average received power to the distance between the transmitter and the receiver, according to the general inverse propagation law $P_r(r) = Ar^{-\beta}$. Typical value for a city environment -where the UMTS will be very popular- are: cell side (max r) = 200m , A = -30dB and $\beta = 3.5$. For these parameters the distance-loss curve is illustrated in the following figure:

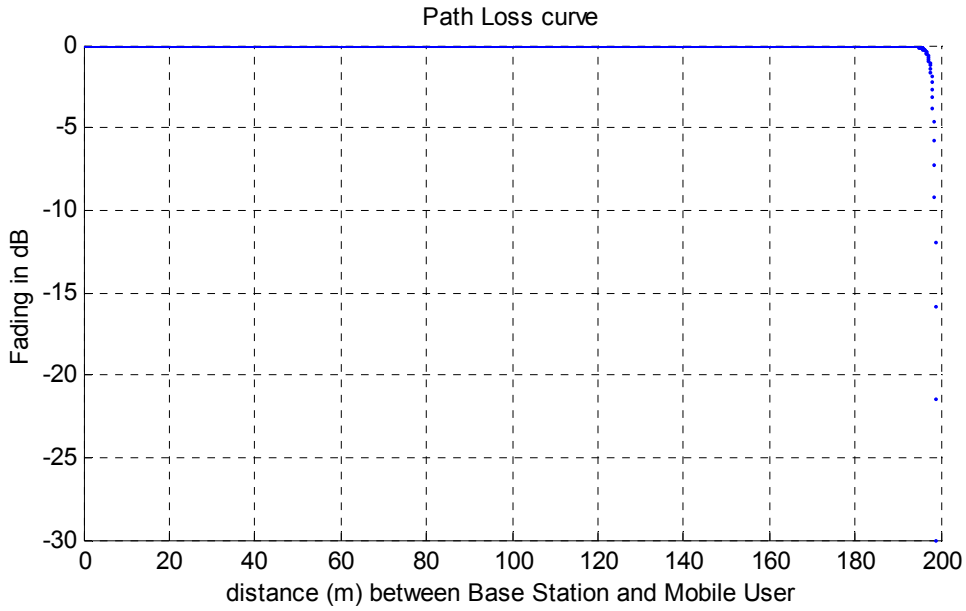


Figure 3. 2: Path Loss Curve

From the above figure is clear that as the user moves towards the edge of the cell the Path loss increases rapidly.

Shadowing is modeled by a multiplicative log-normal random variable with dB-spread $\sigma = 4$dB , i.e. a random variable whose value expressed in dB has Gaussian distribution with zero mean μ and $\sigma = 4$ dB ($F(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{(\log x)^2}{2\sigma^2}}$) . In the next figure we illustrate the shadowing fading distribution using one million samples:

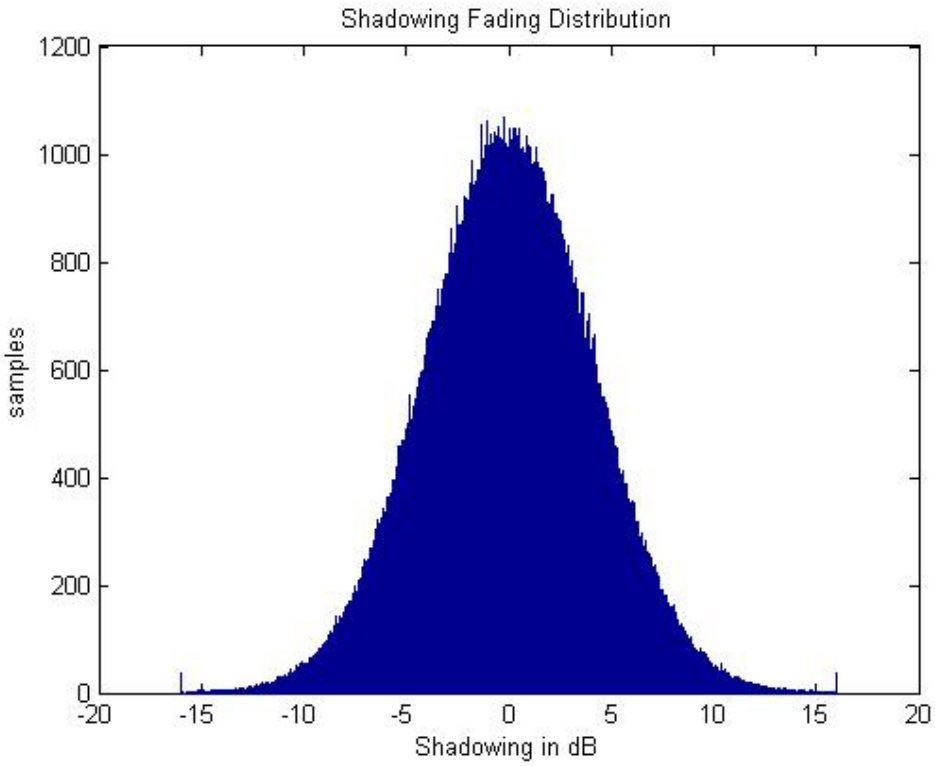


Figure 3. 3 Shadowing Fading distribution

Both Path Loss and Shadowing are dynamically changed during the movement of each user.

3.1.2 Wideband Effect Loss

In a wireless environment mobile users are moving all around. This moving process causes the Rayleigh Fading which is fast for slow moving users and slow for fast moving users. In order to model this fading, a new parameter has been introduced, the Doppler frequency. We assign low Doppler frequencies for slow moving users and higher ones for fast moving users [6]. In the following Table we illustrate some typical Doppler frequency values in a WCDMA environment:

Average Speed of user (Kmph)	Doppler Frequency (Hz)
5	6
17	20
50	40
80	80

Table 3. 3: Typical Doppler Frequencies in WCDMA

Spread spectrum techniques are included in the WCDMA (Direct Sequence-DS (IS-95))[7],[8]. In DS user bits are coded by a unique binary sequence known as the code. The bits of the code are called chips. The chip rate (W) is typically much higher than the bit-rate (R). The signal spreading is achieved by modulating the data modulated signal a second time by using wideband spreading signal. The used code in

WCDMA is a pseudorandom sequence which is constructed by combining Orthogonal Variable Spreading Factor and Cell specific scrambling code. Transmitter (Tx) and Receiver (Rx) are using the same codes which are synchronized, the received narrowband user data is amplified with the factor of W/R (=processing gain). Other power components like interference coming to the receiver will not have processing gain. Processing gain includes spreading gain and channel coding gain. The processing gain is different for different services over 3G mobile network applications (voice, www-browsing, and videophone) due to different R. This means that the coverage area and capacity is different for different services. All the above are illustrated in the next figure [7]:

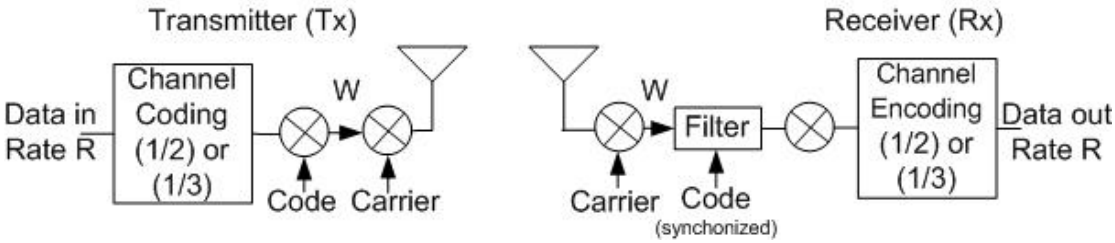


Figure 3. 4: Spread Spectrum Technique in WCDMA

Radio channel consists of many multipaths. Each multipath changes the amplitude and the phase of the transmitted signal. The data in QPSK signal is in phase. Energy splitted into many fingers is detected by the matched filter as shown in Figure 3.5. Maximum Ratio Combining (MRC) corrects the phase with channel amplitude estimation [7].

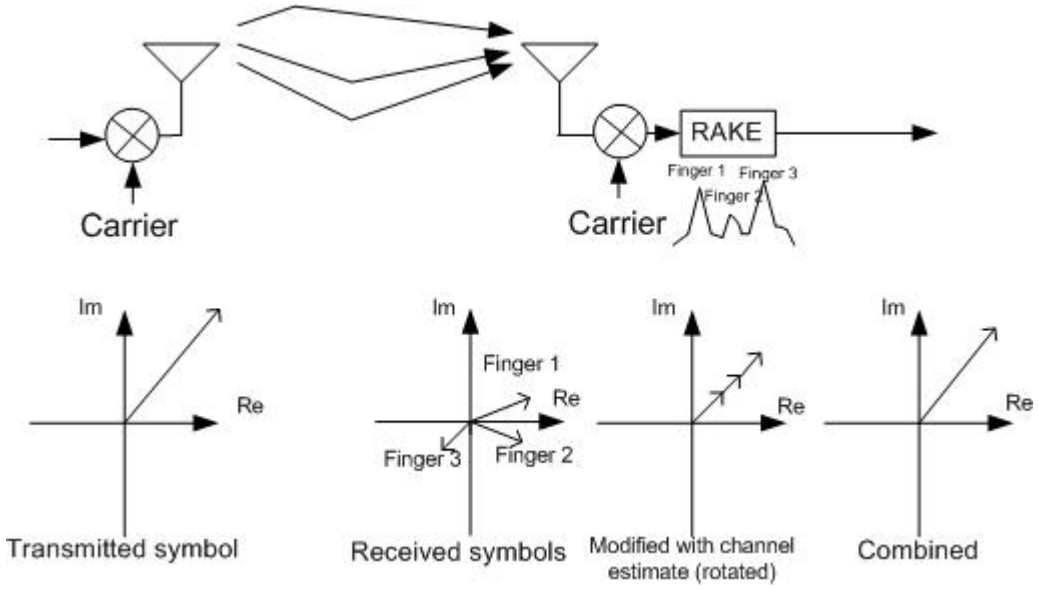


Figure 3. 5: Maximum Ratio Combining in RAKE

The RAKE receiver shown in the previous Figure is known as the RAKE diversity receiver and is illustrated in the following figure.

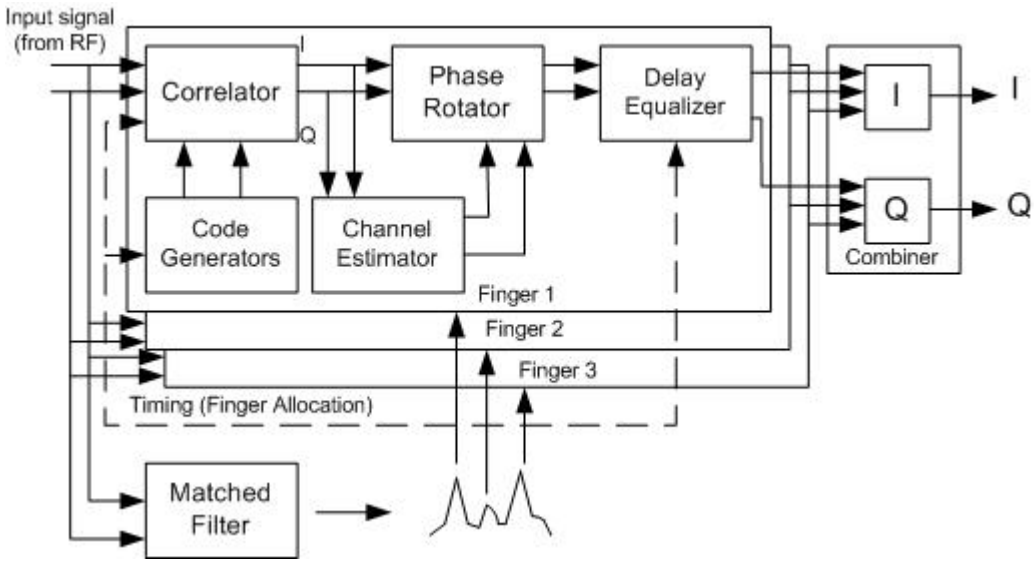


Figure 3. 6: RAKE diversity receiver

In order to despread the incoming signal with the code, the timing has to be known. This can be detected by a matched filter as shown in the following figure. When samples of incoming serial data are equal to bits of predefined data, there is a maximum at filter output.

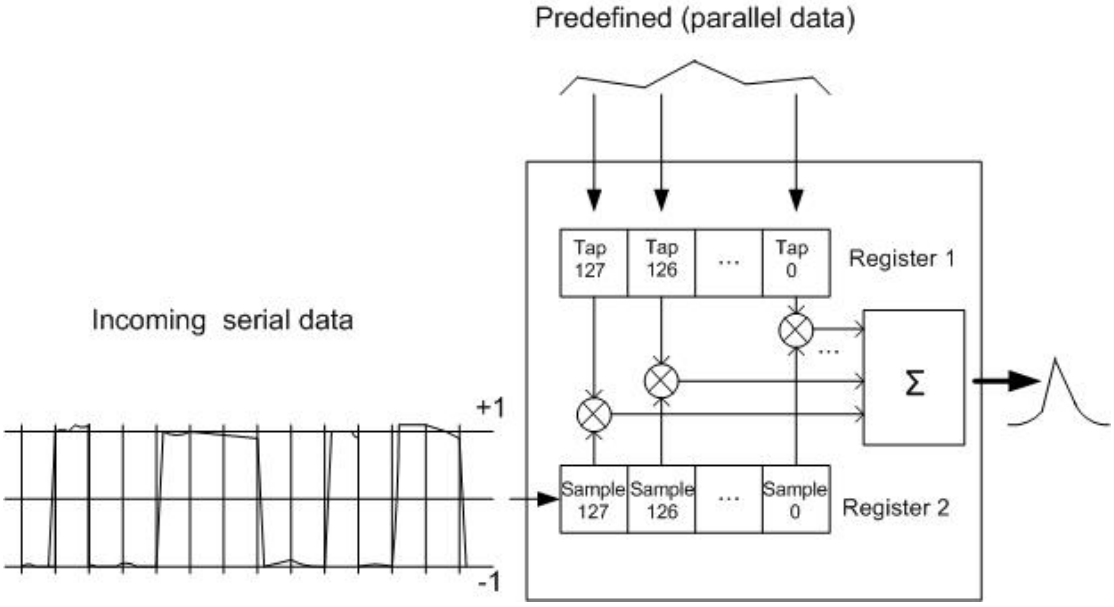


Figure 3.7: Matched Filter

Because of the multipath propagation the received signal is a sum of components with random phase and amplitudes. Depending on the relative phases of these components the sum field is either amplified or attenuated. This causes the fading received signal envelope. In RAKE each receiver tap includes many multipath propagated signal components. Each tap can be then assumed to be independently fading (typically Rayleigh fading). Using the matched filter we can estimate the delay profile of the signal. Multipath propagation causes several peaks in matched filter output and allocates RAKE fingers to these peaks.

In the following Figure we illustrate the WCDMA transmission's block diagram that sums up all the aspects we have presented in this section.

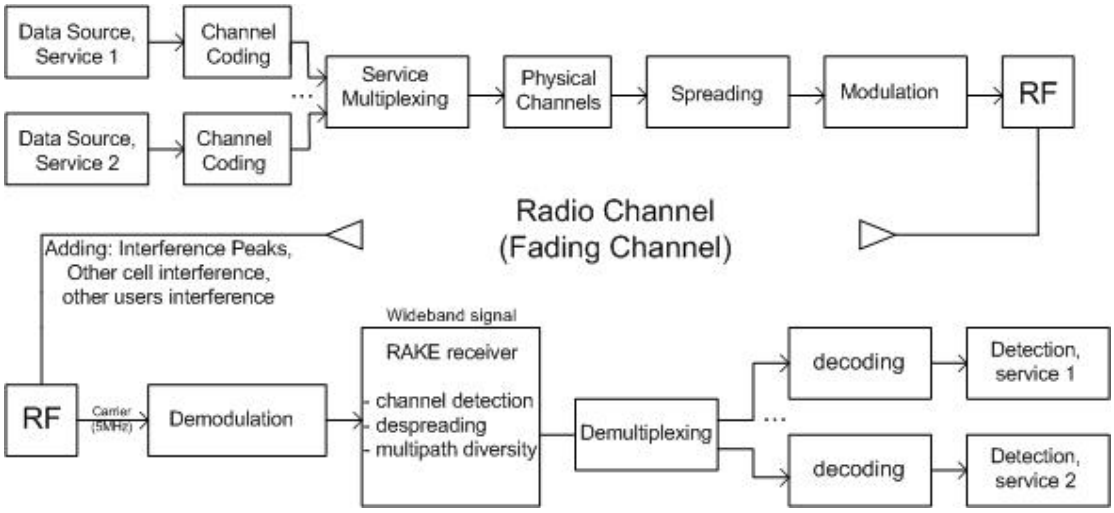


Figure 3. 8: WCDMA Transmission

3.2 Radio Resource Management

The Radio Resource Management (RRM) is a management responsibility solely taken care of by UTRAN. RRM is located in both UE and RNC inside UTRAN. RRM contains various algorithms, which aim to stabilize the radio path enabling it to fulfill the QoS criteria set by the service using the radio path [9].

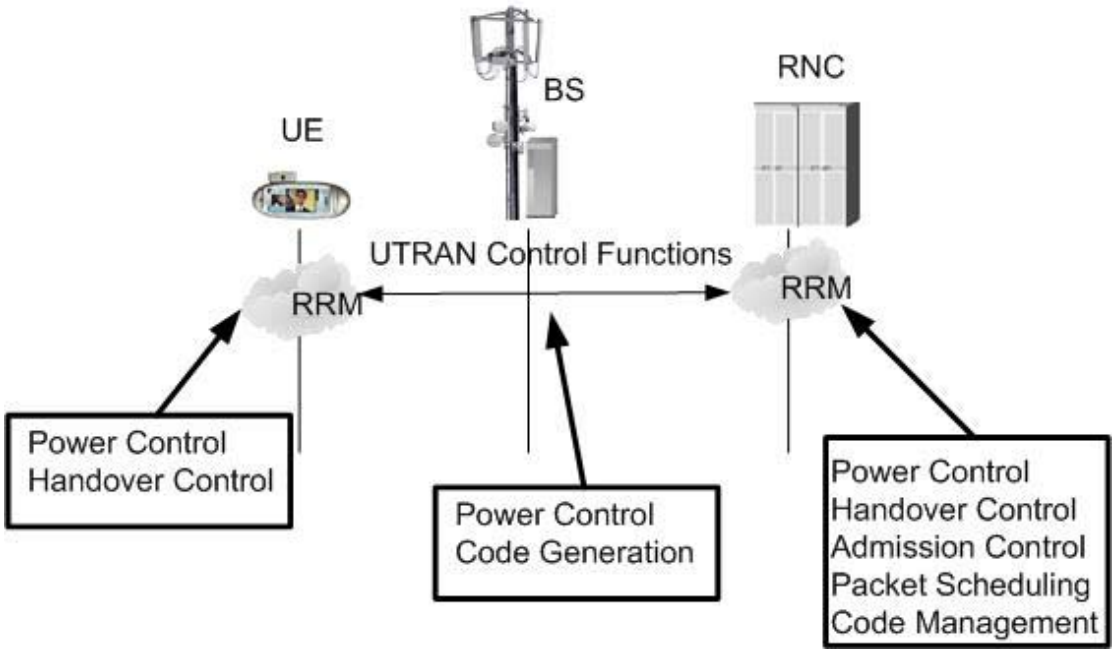


Figure 3. 9: RRM and Radio Resource Control

The RRM algorithms must deliver information over the radio path, which is named UTRA Service. The control protocol used for this purpose is the Radio Resource Control (RRC) protocol. The UTRAN control functionalities are discussed later. The RRM algorithms to be shortly presented here are:

1. Handover Control
2. Power Control
3. Admission Control (AC) and Packet Scheduling
4. Code Management

3.2.1 Handover Control and Macrodiversity

Handover is one of the major headaches not only for the network planning and operating engineers but also for the research and development engineers. It is very difficult and complicated to introduce and use a general purpose handover model due to the heterogeneous topology of the UMTS network. Moreover handover is maybe the main reason for link interrupts. During the handover not only voice but also data can be lost. Furthermore the handover frequency increases as the cell radius decreases (UMTS case study) or speed of mobile users increases.

In UMTS there are two major types of handover: intra-system handovers and inter-system handovers. Intra system handovers are classified into two types: intra-frequency handovers and inter-frequency handovers.

Intra-frequency handovers are classified into three types:

1. Softer where the User Equipment is connected to two sectors of the same Base Station simultaneously (no delays).
2. Soft where the User Equipment is connected to two sectors of different Base Stations simultaneously (no delays).
3. Hard where the User Equipment is connected to only one sector at the time (causing short delays as we will explain later).

Inter-frequency handovers are classified into two types:

1. Where the User Equipment is connected to only one sector at the time working at different frequencies (causing short delays).
2. Between cell layers that is from macrocells to microcells.

Inter-system handovers are handovers between WCDMA and GSM 9000/18000 or between WCDMA and some other system.

3.2.1.1 Soft and Softer Handover

In Soft handover the User Equipment is connected to two or more base stations and in Softer handover is connected to two or more sectors of the same Base Station at the same time. This means that the same information flows through many Base Stations and the receiver has the duty to receive all these signals. The User

Equipment enters in the handover state if the difference between measured pilot signals (measured pilot E_c/I_0) from several Base Stations are within the threshold value.

As shown in the following figure, when the mobile is close to the cell border between BS1 and BS2 and moving towards the cell of BS2: (1) The strength of the signal from BS1 becomes equal to the defined lower threshold. On the other hand, based on the User Equipment measurements the RNC recognises that there is already a neighbouring signal available, which has adequate strength for improving the quality of the connection. Therefore it adds the Signal B to the active set. Upon this event, the User Equipment has two simultaneous connections to the UTRAN and hence it benefits from the summed signal, which consists of signal from BS1 and BS2. (2) At this point the quality of signal from BS2 starts to become better than that from BS1. Therefore the RNC keeps this point as the starting point for the handover margin calculation. (3) The strength of signal from BS2 becomes equal or better than the defined lower threshold. Thus its strength is adequate to satisfy the required QoS from the connection. On the other hand, the strength of the summed signal exceeds the defined upper threshold, causing additional interference to the system. As a result, the RNC deletes the signal from BS1 from the active set [11].

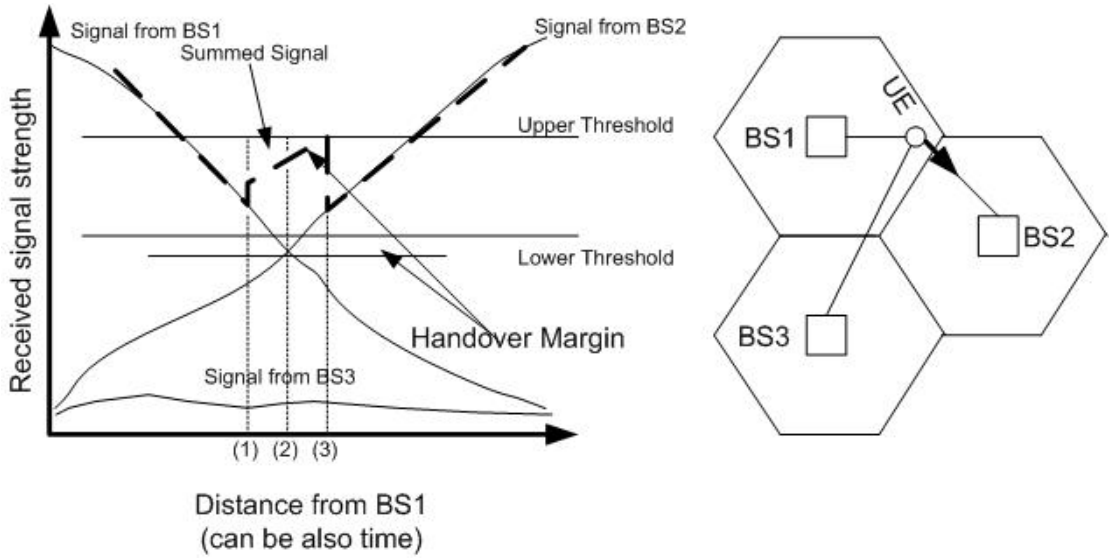


Figure 3.10Soft/Softer Handover Algorithm

It should be noted that the size of the active set may vary but usually it ranges from 1 to 3 signals. Because the direction of the UE motion varies randomly, it is possible that it comes back towards the cell of BS1 instantly after the first handover. This result is a so-called ping-pong effect, which is harmful for the system in terms of capacity and overall performance. The reason using the handover margin is to avoid the undesired handovers, which cause additional signaling load to the UTRAN.

The main reasons for which the UMTS forum has chosen Soft/Softer handoff as the standard handoff technique are the followings: Using this type of handoff the interference, from one sector of the Base Station (Softer handoff) or Base Station (Soft handoff) to another, decreases and therefore the capacity increases. Without Soft/Softer handoff the interference power would be very high because the same frequency is used in adjacent cells. Moreover there is a convergence gain due to diversity. The signal to other Base Stations can be temporarily very weak. Because of

this type of handoff, the resulting signal can be relatively good but on the other hand additional resources at the Base Station and additional signaling is needed. Last but not least during Soft/Softer handover there is no delay as the above handover algorithm uses streaming real-time switching techniques with direct communication between the two RNCs, using the Iur interface, as shown in the following figure [1].

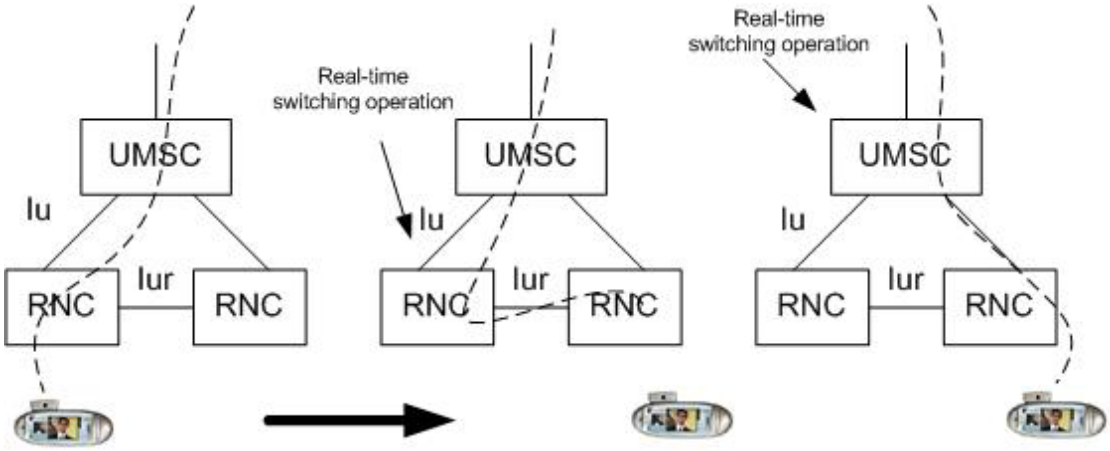


Figure 3. 11: UMTS streaming for RNC changes

3.2.1.2 Other handoffs

If it is not possible for the UE to handover using Soft/Softer handoff it has to use hard handoff. This can happen due to load reason, coverage reason or even service reason. In this case there will be some delay. In rural areas and especially for voice services the operator can even choose the GSM network to handoff.

3.2.1.3 Macrodiversity

Because of the fact that the UE may use cells belonging to different BS's or RNC's the macrodiversity functionality also exists at the RNC level. However, the way of combining the signals is quite different than in the macrodiversity case at the BS due to the RAKE receiver, so as a result someone could claim that macrodiversity at the RNC level would be a solution for faster connections. Therefore, other approaches like the quality of data flow – that is related to the TCP as we will explain – can be utilized to combine or just to select the desired data stream. What is also important is that macrodiversity takes place only at the server side but microdiversity (using RAKE receiver) which is responsible for combining weak signals after following multipath for both uplink and downlink that is at the BS or UE, respectively.

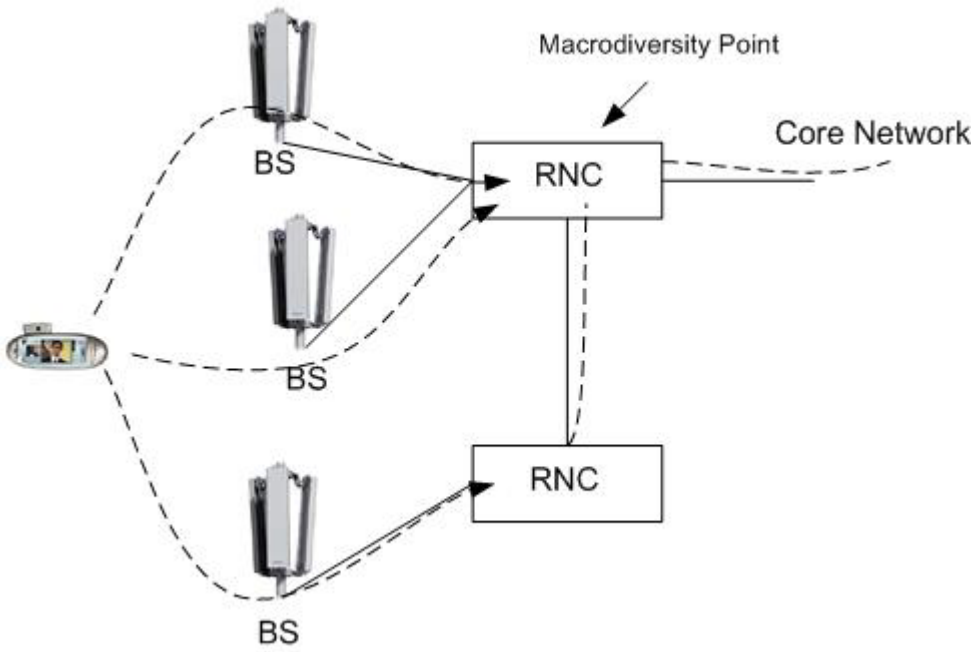


Figure 3.12: Macrodiversity in RNC

3.2.2 Power Control

The purpose of the power control is to ensure that every mobile and base station transmit just enough energy to convey information while interfering other users as little as possible. The ideal would be every user to transmit and receive the same power, but this is impossible not only because each user is located at a different place but also because the received power of one user reduces as the user population in the cell increases. This is illustrated in the following figure (W=3.84 Mcps, R=12.2 Kbps, threshold:-103dBm [dBm = ($10Log_{10}$(milliWatts))]) [7].

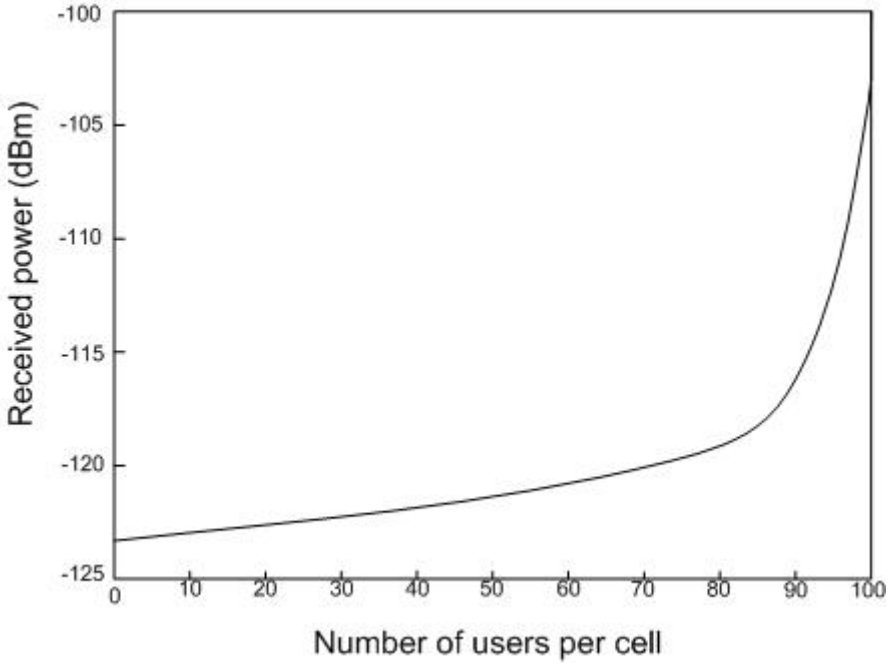


Figure 3. 13: Received power of one user as function of users per cell

There are three different power controls:

1. Open Loop: Open Loop power control that is used for initial power settings of the Mobile Station.

2. Closed Loop: Across the air-interface Closed Loop Transmitter Power Control (TPC) at the rate of 1.5 KHz. This control mitigates fading processes, both fast and slow fading, on uplink and downlink. Moreover uses a fixed quality target (SIR) set in Mobile Station and Base Station.

3. Outer Power: Outer Power control loop that sets the SIR target used by the open loop power control based on the Frame Error Rate(FER) detected at the radio network controller (RNC). Moreover, it compensates changes in the propagation conditions and adjusts the quality target (SIR) set to achieve a target FER/BER both on uplink and downlink. The mobile station speed and the available multipath diversity are input parameters of this power control loop.

Power control also tries to solve the problem of the Near-Far problem in CDMA. Without power control every Mobile Station with full power (Mobile 2) blocks the cell because Mobile Station 1 and Mobile Station 2 use the same frequency and transmit simultaneously. On the other hand with power control everything rules just fine. The next two figures illustrate the problem and the power control solution.

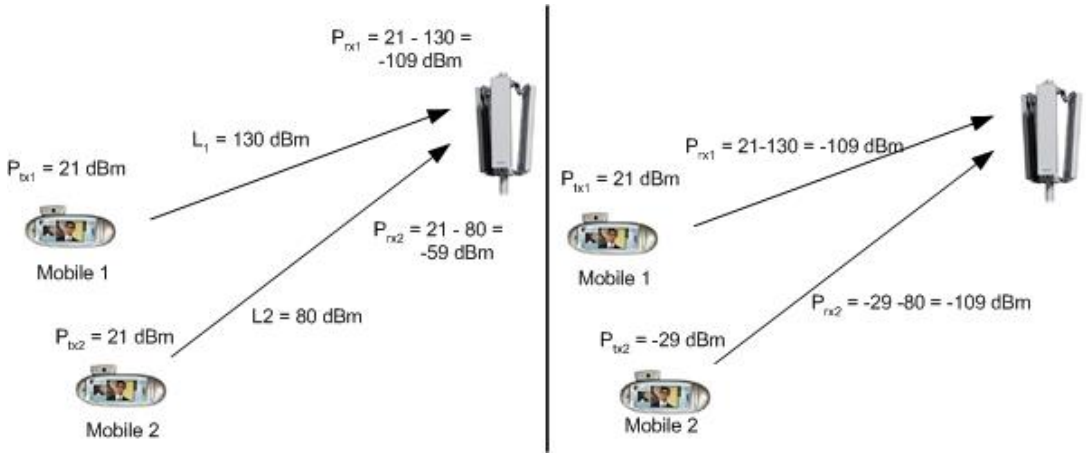


Figure 3. 14: Near-Far Effect without and with power control

The faster the power control the better is the performance. In WCDMA the power control can compensate even the fast fading with low mobile speeds. Fast power control runs also in downlinks. We are now ready to give some more details about the default type of power control, The Closed Loop.

Fast Closed Loop power control can be divided into two categories:

1. Uplink Transmitter Power Control .In this case SIR target is defined at the Base Station. If the measured SIR at Base Station is lower than the SIR target, the Mobile Station increases its transmitted power, in other case it decreases its power. The power control rate is 1500 Hz.

2. Downlink Transmitter Power Control. In this case if the measured SIR at the Mobile Station is lower than the SIR-target coming from the outer loop of the Mobile Station, the Base Station increases its transmitted power for that mobile, else it decreases its power. Power control rate is 1500 Hz and the power control is dependent on the service. There is no near-far problem in the downlink because Transmitted Power Control saves Base Station power resources.

Figure 3. 15: Uplink and Downlink TPC

The closed Loop power control follows the fast fading when the Mobile Station speed is low. In uplink the Base Station asks the Mobile Station to increase or decrease its power depending on its measurements (measured SIR value). The drawback is that the interference to other cells increases because of the large power peaks when the channel is in deep fade. However the net-effect is positive and performs better with fast Power Control. Fast close loop control tries to keep the SIR constant. The constant SIR does not however guarantee the required frame error rate which can be considered as quality criteria of the link.

One alternative power control is the Outer Loop power control. In contrast with the Closed Loop power control, the frame reliability information has to be delivered to Outer Loop control. SIR target can be tuned according to this information. This control loop runs between the network (RNC/BS) and the Mobile Station as shown in the following Figure.

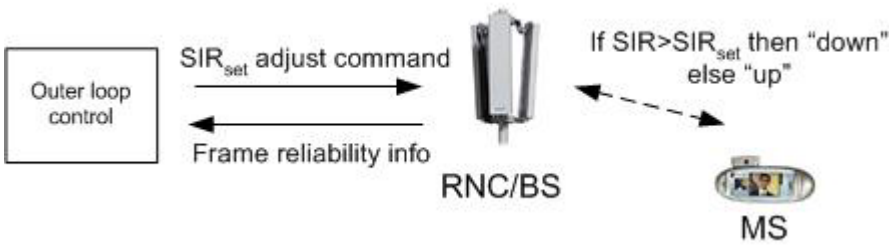


Figure 3. 16: Uplink Outer Loop power Control

3.2.3 Admission Control

Due to the fact that users have different kinds of subscriptions and different QoS requirements, it is important for the system to prioritize users when setting up connections. The admission control function will evaluate requests for new

connections and changes to existing ones in order to ensure that it makes fair decisions. If the system is fully loaded, some users might not be admitted, while high-priority users could be let in, while reducing bandwidth for others who are already connected. This process occurs in accordance with the QoS agreement that is established between the subscriber and the operator.

3.2.4 Code Allocation

In TDMA systems such as GSM, users are separated within a cell by using different time slots (taking turns at sending). In CDMA, everyone sends at the same time, and different codes separate users. These codes are chosen so that the interference between different users and different cells is minimal. The optimal case would be if everyone could use orthogonal codes (codes that do not interfere each other at all). This situation is possible to some extent but requires careful code planning by the system. Most of this code planning is done automatically by the network.

3.3 Network Planning

Admission control, together with power control, ensures that a single MS will not consume the entire power level of one base station. In the uplink there is less flexibility, however, because an MS has a limit to how much power it can use. Usually, this problem is solved by letting the MS go down in bit rate as it moves farther away from the antenna, unless it finds a new cell that is closer to join. This factor is important to consider because an application must be prepared to accept a lower bit rate when this situation happens and gracefully degrade service if possible (like cutting the frame rate of a video). Moreover in order for the rural areas to be covered better, it can be beneficial to use an omnisector (that is a 360^0 coverage like the one in Figure 3.17), while more traffic-dense areas are better suited for a three-sector (that is a 120^0 coverage like the one in Figure 3.18 for microcells) or even better a six-sector base station.

Because the power from the base station in the downlink is also limited, having more users in the cell will give less power to each MS. In effect, the size of the cell becomes smaller because the base station in a fully loaded cell only has power to reach those that are closest. This phenomenon is called Cell Breathing (CB) and is somewhat tricky to handle. Figure 3.18 shows this phenomenon. Every infrastructure vendor must tackle this issue, and the Admission control function can help in that respect.

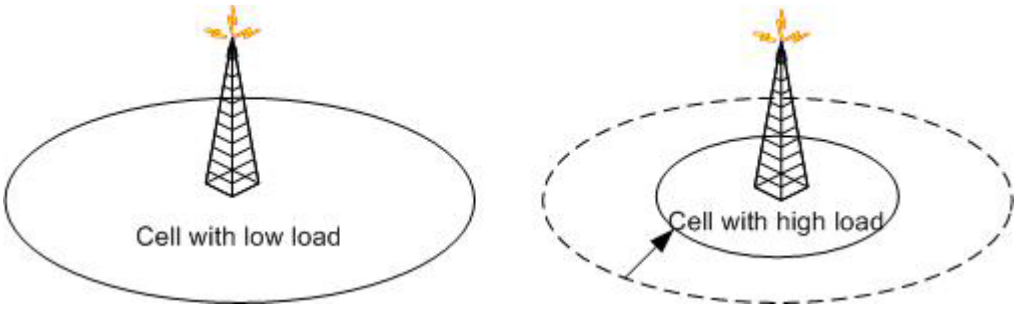


Figure 3. 2: Cell Breathing

Cell breathing also becomes less of an issue with the introduction of hierarchical cells. A large cell can cover the same area together with a number of small cells, acting as the backup for the smaller ones and as the preferred cell choice for users who are moving fast. In Figure 3.18 the mobile user in the car is preferable to be connected to the macro cell (if the transmission protocol that it uses can be adapted to such an environment as we will explain in a forthcoming chapter), or it would have to change micro cells very often imposing a large load on the system.

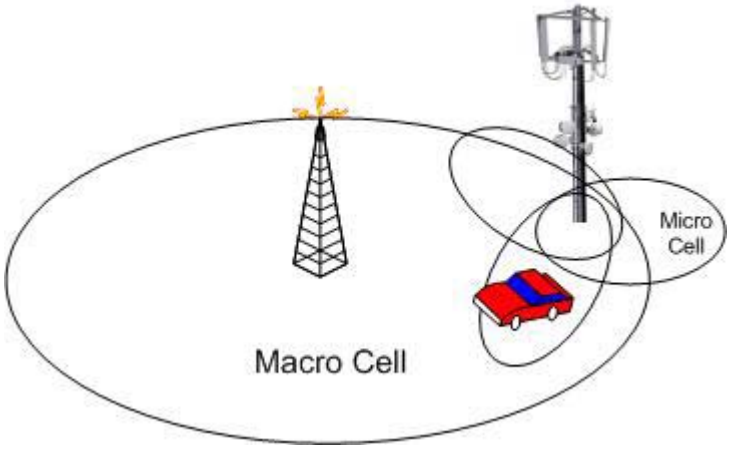


Figure 3. 3: Macro-Micro Cells

3G terminals will be capable of performing many different tasks at the same time, so it is essential to add this functionality in a way that is efficient for both the network and the terminal. The problem however, is that then the terminal needs to be equipped with transceiver equipment for each physical channel (more physical channels may be used), which adds cost to the handsets. That is why 3G CDMA, as mentioned before, enables several logical channels (voice, data, packet switched and circuit switched) over a single physical channel. Last but not least, the core network will not affect the end user and the application in term of losses to the same extend as the radio interface will.

3.4 Cell Capacity

In the following we present the basic idea how to roughly estimate WCDMA Transceiver (TRX) capacity theoretically and based on radio conditions. In order to simplify the issue, one must make some assumptions [11]:

1. All the subscribers under the TRX coverage are equally distributed so that they have equal distances to the TRX antenna.
2. The power level they use is the same and thus the interference they cause is on the same level.
3. Subscribers under the TRX use the same base-band bit rate that is the same symbol rates.

Under these circumstances a value called Processing Gain (G_p) can be defined. PG is a relative indicator giving the relationship between the whole bandwidth available (c) and the base-band bit rate ($B_{Information}$): $G_P = \frac{B_{RF}}{B_{Information}}$

There is another way to express G_P by using the chip and data rates:

$$G_P = \frac{Chiprate}{Datarate}$$

Both ways (when presented in dB values) give as a result the improvement of the Signal to Noise Ratio (S/N) between the received signal and the output of the receiver.

Further on, the processing gain is actually the same as the spreading factor. It should be noted that the base-band bit rate discussed here is the one achieved after the rate matching. In this process the original user bit rate is adjusted to the bearer bit rate. Bearer bit rates are fixed e.g. 30/60/120/240/480/960kbps. The system chip rate is constant: 3.84Mcps. Hence, as an example the bearer having a bit rate 30kbps will have a spreading factor 128:

$$G_P = \frac{3840000}{30000} = 128 = \text{Spreading Factor}$$

The power P required for information transfer in one channel is a multiple of the energy used per bit and the base-band rate.

$$P = E_b \cdot \text{Baseband Datarate}$$

On the other hand, it is known that the noise on the channel ($N_{Channel}$) using partially the whole bandwidth B_{RF} can be expressed as:

$$N_{Channel} = B_{RF} \cdot N_o$$

Where N_o is the Noise Spectral Density (W/Hz).
Based on this, the signal to noise ratio is:

$$S/N = \frac{P}{N_{Channel}} = \frac{\text{Eb} \cdot \text{Baseband Datarate}}{B_{RF} \cdot No} = \frac{E_b / No}{G_P}$$

If assumed that there are X users under the TRX and the assumptions presented earlier are applied, it means that there are (statistically thinking) X-1 users causing interference to one other. This is also indicates signal to noise ratio and then expressed in mathematical format the outcome is the following equation:

$$S/N = \frac{P}{P \cdot (X-1)} = \frac{1}{X-1}$$

If, further on, there are plenty of users (tens of them) then the equation could be simplified:

$$S/N = \frac{1}{X-1} \approx \frac{1}{X}$$

Now there are two different ways to calculate signal to noise ratio:

$$\frac{E_b / No}{G_P} \approx \frac{1}{X} \Rightarrow \mathrm{X} \approx \frac{G_P}{E_b / No}$$

This equation is very rough expression and should be used for the termination purposes only. The "official" way to calculate TRX capacity has several more parameters to be taken into account.

Assuming that the spreading factor used in the cell is 128 and that for the transactions used the E_b/N_o is 3dB. If there is one TRX the number of users that one cell can contain simultaneously is:

$$\mathrm{X} \approx \frac{G_P}{E_b / No} = \frac{128}{3dB} \approx \frac{128}{2} = 64 \text{ users}$$

This is the maximum number of users the TRX is able to handle in theory, when taking into account the intra-cell interference. In reality the neighboring cells produce inter-cell interference. If assumed that the inter-cell interference is the same as intra-cell interference then the user number reduced by a factor of two is 32 (=64/2).

The E_b/N_o relationship is the point of interest, is a constant-like numerical relationship which may have several values, which later are related to radio interface bearer bit rates. Thus it can be stated that the E_b/N_o relationship has a remarkable effect on the TRX/cell capacity as far as the maximum number of simultaneous users is concerned. In the E_b/N_o relationship the following issues should be considered:

1. N_o is a local constant value type, which also contains some receiver specific values
2. E_b is a changing value in nature and its dependencies are described in the following points
3. Processing gain/spreading factor: the bigger the spreading factor value, the smaller the E_b.
4. The higher base-band bit rate used, the bigger the E_b will be (this is a direct consequence from the previous point).
5. Distance between terminal and BS receiver: the longer the distance, the bigger E_b.
6. Terminal motion speed: the higher the speed used the bigger the E_b.

The calculation above is a rough way to estimate TRX capacity; there are many other ways to do this.

CHAPTER 4: UMTS Core Network

4.0 Introduction

The UMTS Core Network can be seen as the basic platform for all communication services provided to the UMTS subscribers. The basic communication services include switching of circuit-switched calls and routing of packet data. Value-added services on top of these basic services are included in next chapter.

The CN maps the end-to-end Quality of Service requirements to the UMTS bearer service. When inter-connecting to the other networks the QoS requirements also need to be mapped onto the available external bearer service. This gateway role of the UMTS CN in creation of the end-to-end service is illustrated in the following figure [11]. The external bearer (e.g. the Internet) is not in the scope of UMTS system specifications and this may create some local problems in the QoS requirements to be satisfied between the UMTS and the external network.

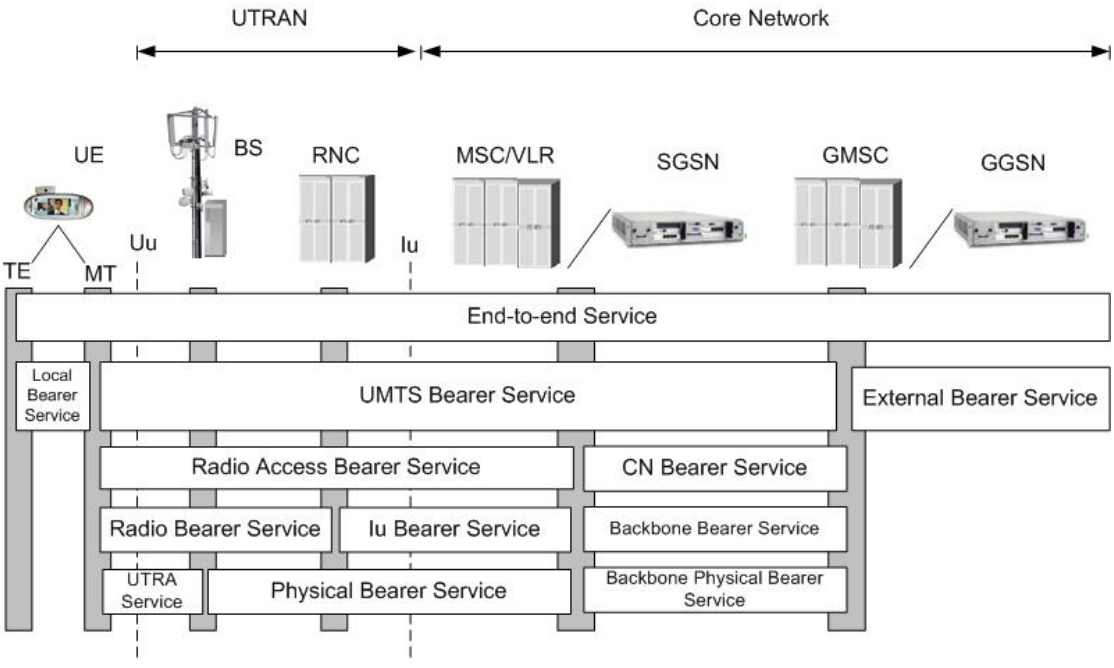


Figure 4. 1: Bearer and QoS architecture in CN

Between the MT and CN the QoS is provided by the radio access bearer hides the QoS handling over the radio path from the CN. Within the CN the QoS requirements are mapped to its own bearer service, which in turn is carried on the backbone bearers on top of the underlying physical bearer service. A challenge in the CN implementation is that the operator has pretty much freedom in choosing how to

implement the physical backbone bearers. The physical backbone bearers rely on the physical transmission technologies used between the CN nodes. Typical transmission technologies like Plesiochronous Digital Hierarchy (PDH) and Synchronous Digital Hierarchy (SDH) with Pulse Code Modulation (PCM) channeling or with the ATM cell-switching are selected.

4.1 Core Network Architecture

The Core Network contains two separate domains for traffic delivery and these domains take into account the special characteristics of the traffic. The special characteristics of the traffic also affect the CS and PS domain element addressing scenarios and further on the signaling interfaces and their transport. The CS domain uses GSM inherited signaling scenarios based on a Mobile Application Part (MAP) protocol covering any possible add-ins the UMTS brings into the system. In the following figure these inherited interfaces are marked with capitals according to the MAP interface naming rules. These interfaces follow the same functioning principles as those already used in the GSM system [11].

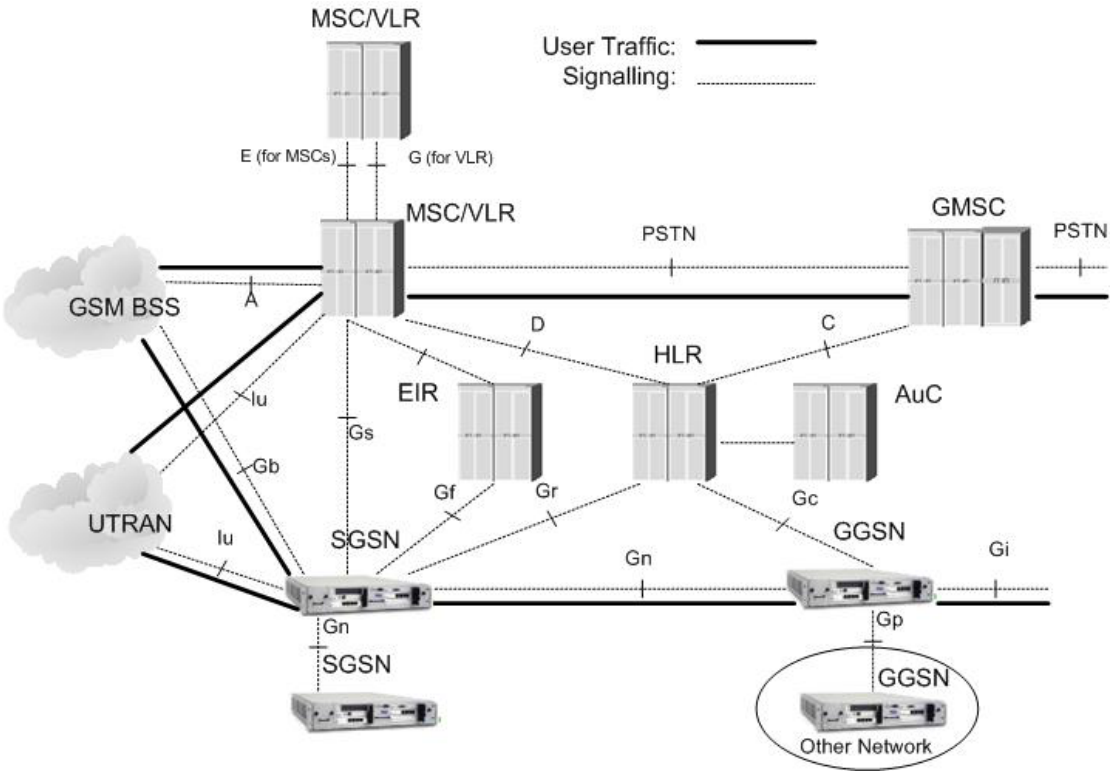


Figure 4. 2: 3GPP R99 CN interfaces

The PS domain in UMTS is evolved from the 2G GPRS and this can be seen from the inherited interface names always starting with G and having another letter indicating which interface is in question. From the functional point of view the Gx interfaces resembles their CS domain counterparts. For instance, the interface Gc uses similar procedures and mostly the same parameters as the MAP interface C; both these interfaces retrieve location information from the HLR.

As already mentioned, in 3GPP R99, the major changes were targeted to the access part of the network; the main new issue presented was new wideband radio access, UTRAN. On the CN side the aim was to minimize changes and utilize the existing GSM/GPRS network elements and functionalities as much as possible. In 3GPP R4 the strategy is somewhat reversed: the access network does not experience much change but the CN is extended remarkably.

The MSC/VLR evolves into MSC server and MG. The MSC server is a network element containing CM main functionality that is to maintain logical Communication Management. MSC server is also responsible for Mobile Management, as we will show in the next paragraphs, and MSC server also contains VLR. The MGW contains the facilities to perform actual switching and network internetworking functions.

The division between MSC server and MGW is not one-to-one; one MSC server may control numerous MGWs, this brings scalability into the system. On the other hand, one must take redundancy and security into account when planning the CN CS domain; if the number of MSC servers is under dimensioned, the network may easily suffer from outages and relatively big numbers of subscribers will not gain circuit switched services.

In 3GPP R5 the access network will experience more changes and the changes in the CN will be minor in nature. The main issues in 3GPP R5 are GSM/EDGE RAN (GERAN) and IP transport within the access network. In 3GPP R5 the traffic is always packet switched; here the question is whether it is real-time or non-real-time.

The reference architecture for 3GPP R4 and R5 is the same. In the development of R5 the focus has shifted to the PS domain, which has been extended with the IMS functionality. A more structured view of this part of the R5 system with separate layers for user data transport, network control aspects service capabilities is given in the next figure [11]:

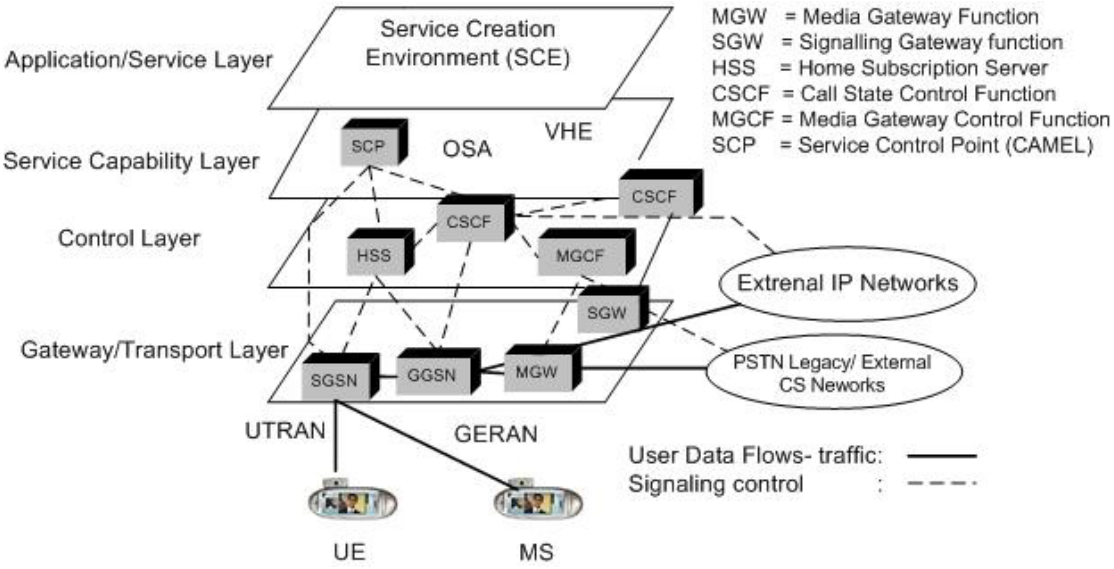


Figure 4. 3: Layered view of the 3GPP R4/R5 CN PS

This service capability layer has already been introduced in 3GPP R99 but in further implementations its role will be increased by Open Service Architecture (OSA) based solutions. The OSA acts as a gateway to the fourth layer of the model, presented in the previous Figure, providing mechanisms for universal service creation and management. Service capabilities are explained in detail in the next chapter.

The changes performed between R4 and R5 should not be visible to the end users. The UTRAN radio path still works like it was working until now; also the terminals being used are still working as such. Within the access network the transport technology could be IP instead of ATM but this is operator's choice.

Besides UTRAN, the evolved GSM BSS named GERAN can be connected to the CN via the Iu interface. Thus traffic coming from GERAN gets the same treatment as the traffic coming from UTRAN with regard to interfaces. If the operator has IMS in use, the CN CS domain is not basically needed any more; the main step (besides the new radio access alternatives) between 3GPP R4/R5 is whether to quit CN CS domain or not. This will depend significantly on the direction and maturing of the VoIP development and penetration.

4.2 Communication and Mobility Management

The main tasks of Communication Management (CM) are communication management and session management. The connection management is a management task responsible for circuit switched transactions and related issues. The control protocols carrying CM information, which deals with call and session control, are related here as a set of Communication Control (COMC) protocols as shown in the following figure.

Te Mobile Management task covers management of UE location together with their identities and addresses, related issues and also security is considered as part of MM. The control protocols supporting execution of MM tasks are referred to as Mobility Control (MOBC) protocols.

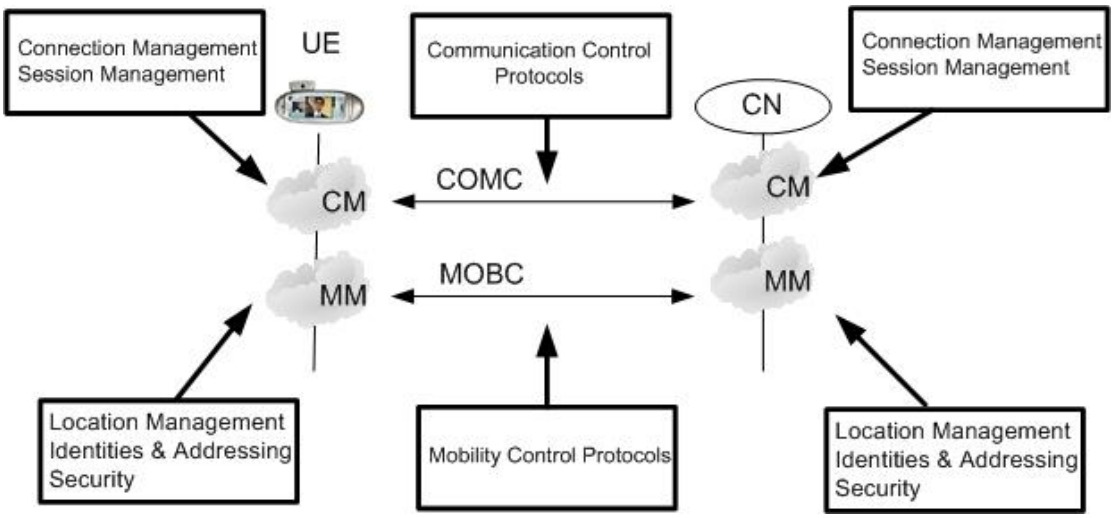


Figure 4. 4: CN Management Tasks and Control Duties

Even in the GSM network, mobility is essential. Understanding the essence of mobility makes the mobile network design significantly different – though more complex as well to the end user. The two basic concepts related to the users' mobility is Location and Position.

The term Location refers to the location of the end user's terminal within the logical structure of the network. The identifiable elements within such a logical

structure are the cells and areas composed of groups of cells. Please note that the word "area" does not refer to a set of geographically neighboring cells but depends on the network operator purposes.

The term Position on the other hand refers to the geographical position of the end user's terminal within the coverage area of the network. The geographical position is given as a pair of standardized co-ordinates. In the most elementary case, when no geographical position can be determined the position may be given as a cell identity, from which the position can be derived.

Although both location and position answer the question, where is this user, the answers are used by the UMTS network in a completely different manner. The Location information is used by the network itself to reach the users whenever there is communication service activity addressed to them. The Position information is determined by the UMTS network when requested by some external service. Although the Position information may well be life-critical (e.g. Emergency call) to the end user, the Location information is life-critical to the network itself in being able to provide services to the mobile users in a non-interrupted manner.

It should be noted that the primary purpose of positioning is to support application oriented services, Position information could also be utilized internally by the network. Examples of these applications are position-aided handover and network planning optimization.

Another key service created by the MM is roaming. The MM functions inside a single Public Land Mobile Network (PLMN) allow a UMTS user to move freely within the coverage are of that single PLMN. Roaming is a capability, which makes it possible for the users to move also from one PLMN to another operated by a different operator company and possibly even in a different country.

4.3 Classes and Applications in UMTS

UMTS QoS classes, also known as traffic classes, are defined keeping in mind that the classification must be simple. The following four leading principles in this respect can be applied. At first the QoS classes must allow efficient use of radio capacity. The CN and UTRAN must be able to evolve independently. UMTS network must be able to evolve independently from its surroundings networks. On the other hand, the backward compatibility mechanisms must be present. Last but not least, the operator must be able to utilize existing transmission technology within UMTS system in a cost effective way.

From the end user point of view, the impression quality is often related to the delay experienced on the connection. Due to this, the connection delay is the main separating attribute between the UMTS QoS classes. The main UMTS QoS classes are:

1. Conversational class: minimum fixed delay, no buffering, symmetric traffic, guaranteed bit rate.
2. Streaming class: minimum variable delay, buffering allowed, asymmetric traffic, guaranteed bit rate.
3. Interactive class: moderate variable delay, buffering allowed, asymmetric traffic, no guaranteed bit rate.
4. Background class: big variable delay, buffering allowed, asymmetric traffic, no guaranteed bit rate.

The conversional class is the most demanding and then the next three are following in decreasing order. In the table we illustrate the attributes for both UMTS bearer service and UMTS radio access bearer service [11].

UMTS bearer service attributes

	Conversional	Streaming	Interactive	Background
Maximum bitrate(kbps)	Less than 2048	Less than 2048	Less than 2048	Less than 2048
Guaranteed bitrate(kbps)	Less than 2048	Less than 2048	N/A	N/A
Symmetry	Symmetric	Assymetric	Assymetric	Assymetric
Transfer delay(ms)	100-250	250	N/A	N/A

UMTS radio access bearer service attributes

	Conversional	Streaming	Interactive	Background
Maximum bitrate(kbps)	Less than 2048	Less than 2048	Less than 2048	Less than 2048
Guaranteed bitrate(kbps)	Less than 2048	Less than 2048	N/A	N/A
Symmetry	Symmetric	Assymetric	Assymetric	Assymetric
Transfer delay(ms)	80-250	250	N/A	N/A

Table 4. 1: UMTS bearer service attributes

At the CN side, end-to-end service requirements are finally mapped onto the UMTS bearer based on the defined QoS attributes. This mapping is performed by the communication management task in the CN. The following figure illustrates the end-to-end principles.

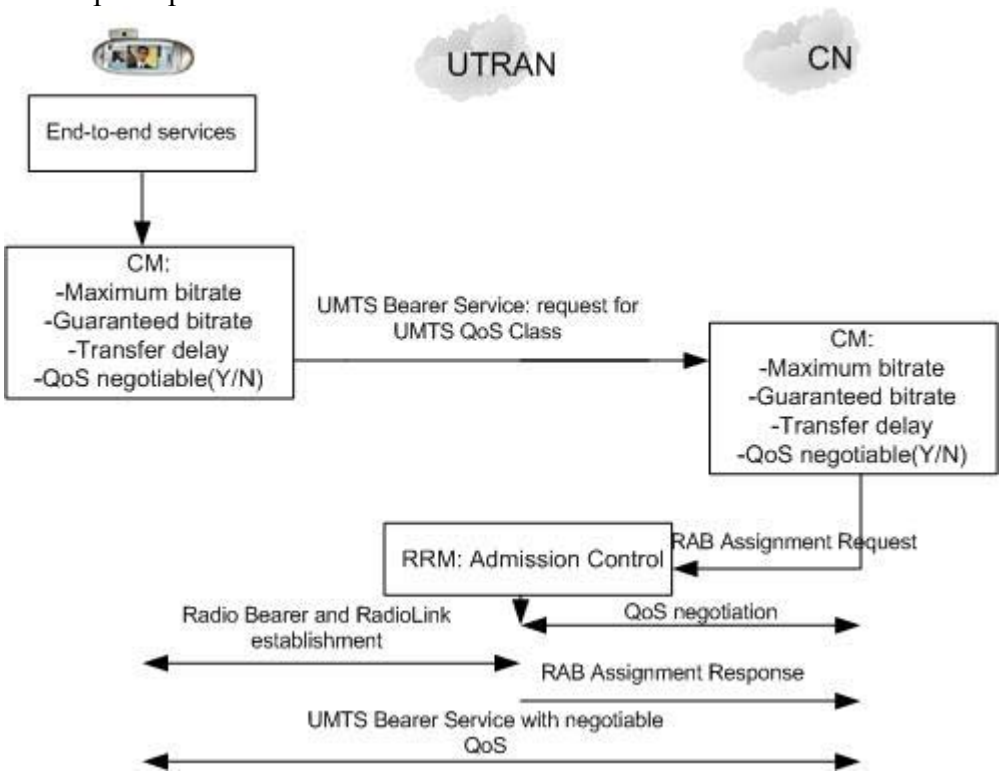


Figure 4. 5: Rough principle of bearer management

The related CN domain also checks the UMTS bearer requirements and starts Radio Access Bearer (RAB) allocation through UTRAN. RAB assignment request is

investigated by the Radio Resource Management (RRM) Admission Control (AC) algorithm, which checks whether the RB for this transaction can be established with the requested QoS parameters. If RBs are not available and the requested QoS is non-negotiable the QoS parameters are renegotiated between UTRAN and CN. If AC in the Radio Network Controller (RNC) allows, the RB is established with the given QoS parameters and then the Iu bearer between UTRAN and CN is also set up. Now the UMTS bearer is ready to carry data flows to the end-to-end QoS requirements (the QoS information is stored in Packet Data Protocol (PDP) context. PDP is a protocol for Radio Management in UMTS).

If a service use conversational QoS class it does not transfer any files from one end to another literally. The first QoS handling file transfer through the connection is the streaming class. Because complete file download takes time and thus causes delay, there must be mechanisms to open and handle files when there are not completely transferred from the source to the destination. This is what the streaming QoS class covers. Typical services or applications using streaming class are those handling big files but showing/playing/handling a limited part of it in time. Also the services offering a multicast (one sender and many receivers simultaneously) type of service utilize streaming class if delay is not an issue, To minimize the possible delay effects the streaming class services are mostly unidirectional; delay exists but it does not cause any harm because the interactivity is missing.

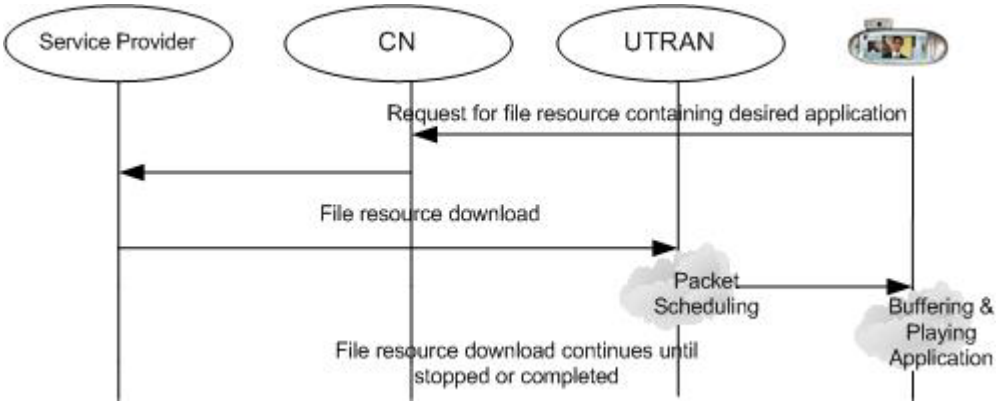


Figure 4. 6: Example of Streaming resource use

CHAPTER 5: The Internet

5.0 Introduction

In order to understand the IP functionality of UMTS, in this chapter, we focus on the principles and characteristics of Mobile Internet - UMTS is also known as an alternative telecommunication system for providing Mobile Internet - and especially on the Transmission Control Protocol (TCP), which is responsible for the internet evolution. The reason we concentrate on TCP is that most of the internet applications use this protocol. We will not discus in extend the use of TCP on the core network of UMTS (which is similar to the current internet infrastructure) but instead on the UTRAN. At the end of this chapter we will be ready to understand why unwiring the internet is a big technology headache.

5.1 Open System Interconnect

In order to understand better the networks, the Open System Interconnect (OSI) model has been introduced. The special OSI which illustrates the Internet is the OSI model for the Internet. The different layers of the Internet model are as follows [3]:

Network Interface Layer: This layer is where the actual bits are transported and the hardware addresses for physical host computers are specified. This layer formats packets and sends them via the underlying network. For mobile users this layer includes the air interface (or ATM part in CN).

Internet Layer: This layer is equivalent to the network layer in the OSI model and includes IP .IP addresses make it possible to locate the destination host and to send packets to it without having to be on the same subnet using Domain Name Server (DNS).

Transport Layer: This layer is responsible to deliver reliably each packet to the appropriate application. Here we find the TCP and UDP versions of the transport potocol.

Application Layer: This layer is responsible for formatting the content and delivering each packet. Here we find Hypertext Transfer Protocol (HTTP) which is a TCP-reliable protocol handling the transfer of Web pages, File Transfer Protocol (FTP) which is a connection-oriented file transfer protocol between two hosts based on TCP and other applications such as multimedia streaming applications.

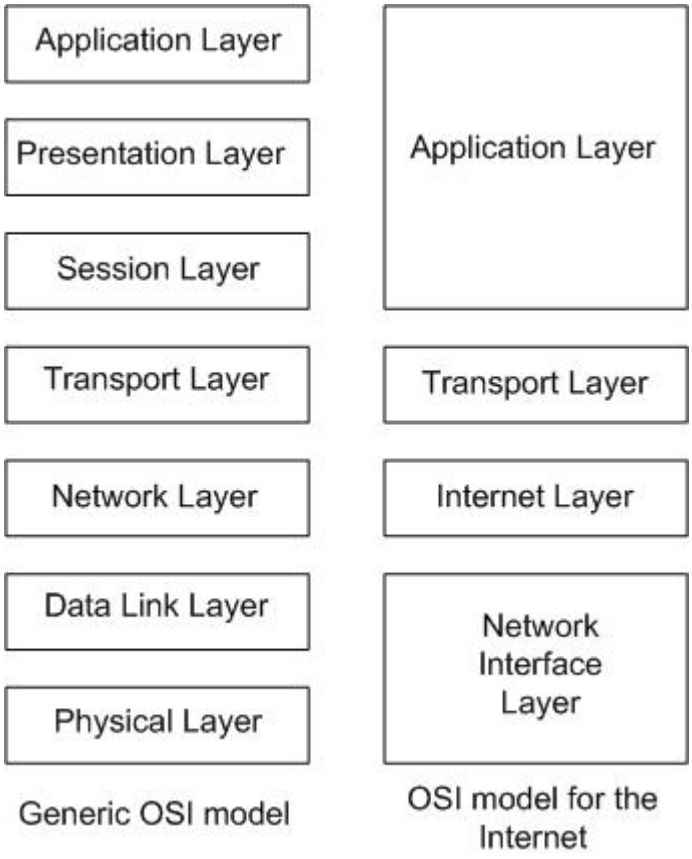


Figure 5. 1: The Generic OSI model and the OSI model for the Internet

5.2 Internet Protocol

The Internet Protocol (IP) transports packets to the desired destination host on the network. IP accepts packets adds its own header and delivers a "datagram" to the data link layer protocol. It is a connectionless protocol and is not aware of any sessions. Every packet is routed independently, and different parts of the same transmission might take a different route. Along the way a packet might be lost, corrupted, duplicated or delivered out of the sequence. If the underlying network is not capable of transmitting packets as big as those that higher layers try to get IP to send, IP will fragment the packets in order to fit the maximum packet size supported by the network. The maximum unit that a packet can be fragmented is called Maximum Transfer Unit (MTU). Incidentally, IP sends packets by using the best-effort principle, and whatever gets lost or received out of the sequence is the responsibility of the higher-layer protocols, to be taken care.

Although the fact that IP was originally designed for wired systems three decades ago, is still THE solution for wired, unwired networks or hybrid networks. There are two versions of IP. IP version 4 (IPv4) is in use and addresses each internet active machine with a 32-bit address. The other version is the forthcoming IP version 6 (IPv6) which addresses each internet active machine with a 64-bit address. The

second version will be the future solution to address the rapidly increasing number of any kind of internet nodes all over the world.

5.3 Asynchronous Transfer Mode

Asynchronous Transfer Mode (ATM) is a network technology (network interface layer) for both local and wide area networks (LANs and WANs) that supports real-time voice and video as well as data. The topology uses switches that establish a logical circuit from end to end, which guarantees the required quality of service (QoS). However, unlike telephone switches that dedicate circuits end to end, unused bandwidth in ATM's logical circuits can be appropriated when needed. ATM is widely used as a backbone technology in carrier networks and large enterprises, but never became popular as a local network (LAN) technology. ATM works by transmitting all traffic as fixed-length, 53-byte (5bytes header, 48bytes load) packets called cells. This fixed unit allows very fast switches to be built, because it is much faster to process a known packet size than to figure out the beginning and end of variable length packets. The small ATM packet also ensures that voice and video can be inserted into the stream often enough for real-time transmission.

The ability to specify a quality of service is one of ATM's most important features. There are four service classes of ATM: Constant Bit Rate (CBR) for real-time voice and video, Real-time variable Bit Rate (rt-VBR) for interactive multimedia, Available Bit Rate (ABR) for data traffic and Unspecified Bit Rate (UBR) for best effort transfers.

5.4 Transmission Control Protocol

The Transmission Control Protocol (TCP) is THE transmission protocol of the internet. The main characteristics of TCP are:

1. TCP is a point-to-point protocol. This means that there is always one sender and one receiver. It creates reliable and in-order in byte stream. Multicasting is not possible with TCP as is.

2. TCP is pipelined. This means that in TCP there is congestion and a window size is set.

3. In TCP there are send and receive buffers as shown in the next Figure [12].

4. TCP uses full duplex communication. There exists bi-directional data flow in the same connection. Moreover there is a Maximum Segment Size (MSS).

5. TCP is connection-oriented. There exist handshaking that is exchange of control messages, initiating sender and receiver state before data exchange.

6. TCP is flow controlled, which means that sender will not overwhelm receiver.

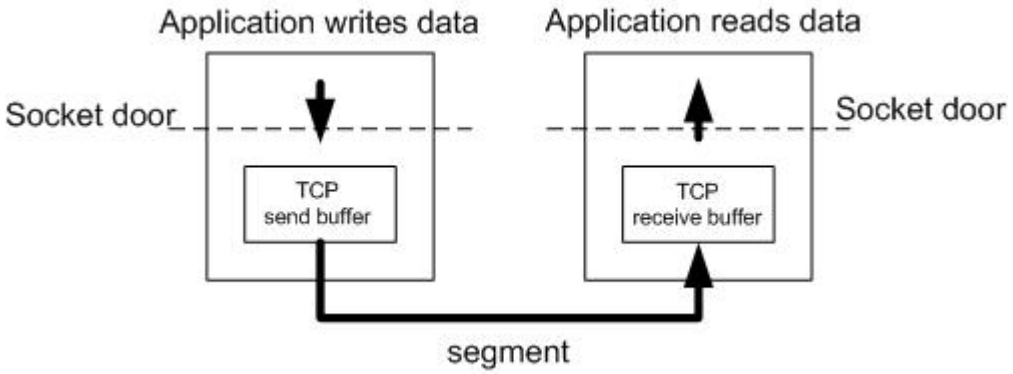


Figure 5.2: TCP send and receive buffers

The header of the TCP segment is 20 byte long. Its major components are: a 2 byte source port number, a 2 byte destination port number, a 4 byte sequence number, a 4 byte acknowledge number, a 2 byte receive window size and a 2 byte checksum.

5.4.1 TCP Reliable data transfer and Retransmission

In TCP the two processes that want to communicate must first handshake with each other. That is why a tree-way handshaking is used. In this type of handshaking, the client first sends a special TCP segment, the server responds with a second special TCP segment and finally the client responds again with a third special segment. The first two segments contain no payload, while the third one may carry payload. More details are provided in the next paragraph.

The Client process passes data through the socket. TCP directs this data to the connection's send buffer (set aside during the initial three-way handshake). From time-to-time, TCP "grabs" chunks of data from the send buffer. Maximum amount of data grabbed and placed in a segment is limited by the MSS. TCP encapsulates each chunk of client data with the TCP header, forming TCP segments. TCP receives a segment, the segment's data is placed in the TCP connection receive buffer. The application reads the stream of data from this buffer. A TCP connection consists of buffers, variables and a socket connection to a process in one host and another set of buffers, variables and a socket connection to a process in another host. No buffers or variables are allocated to the connection in the network elements (e.g. routers) between two hosts. Moreover if a critical time (timeout) has passed without an acknowledgement then a new request is sent.

TCP views data as unstructured but ordered stream of bytes. Sequence number for a segment is the byte-stream number of the first byte in the segment. The acknowledgement number that host A puts in its segment is the sequence number of the next byte host A is expecting from host B. TCP only acknowledges bytes up to the first missing byte in the stream, (i.e. it provides cumulative acknowledgements). When a host it receives out-of-order segments in a TCP connection either discards out-of-order bytes or keeps out-of-order bytes and waits for the missing bytes to fill in the gaps.

Next figure illustrates the two retransmission scenarios of TCP. Host A can be the UE and host B the Base Station or vice versa.

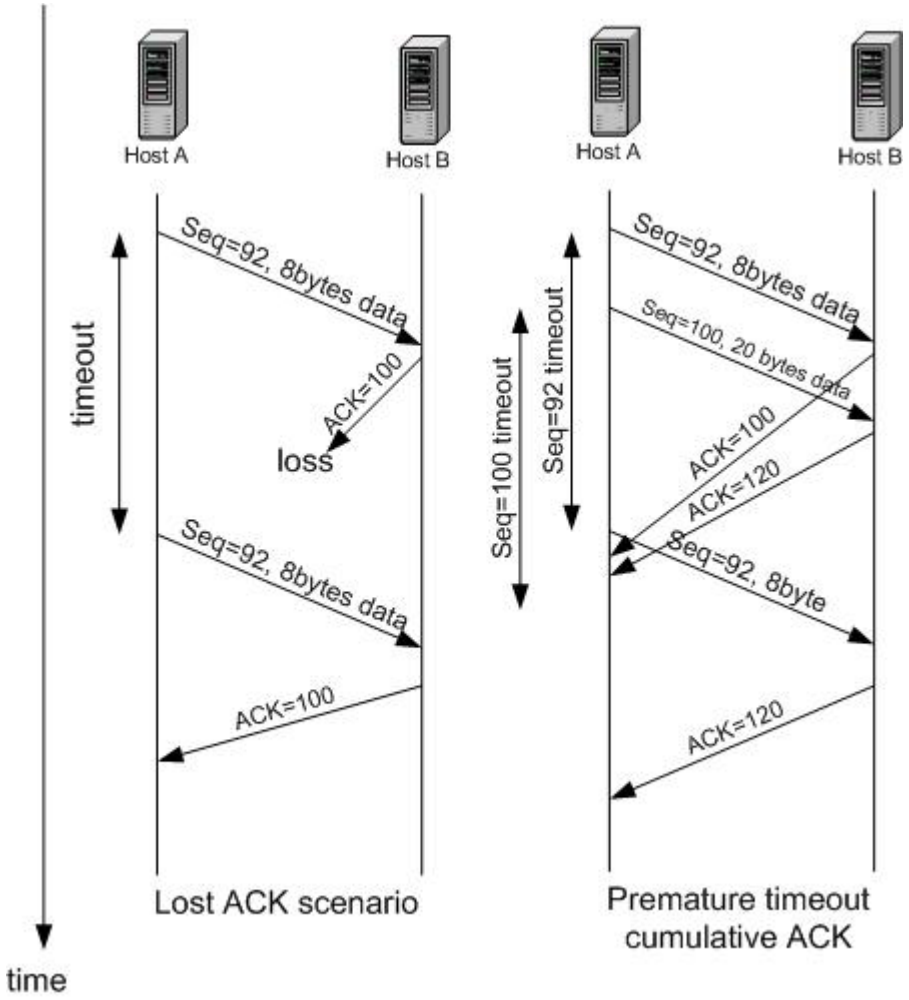


Figure 5.3: Retransmission Scenarios

5.4.2 TCP Flow Control

TCP provides a flow control to its application in order to eliminate the possibility of the sender to overflow the receiver's buffer. This can happen if the sender transmits too much, too fast. In the TCP flow control terminology, RcvBuffer is the size of TCP Receive Buffer and RcvWindow is the amount of space room in the buffer. Using the flow control, the receiver explicitly informs sender of the, dynamically changing, amount of free buffer space (RcvWindow field in TCP segment). On the other hand, sender keeps the amount of transmitted, unAcked data less than the most recently received RcvWindow. The receiver buffering is shown in the next Figure:

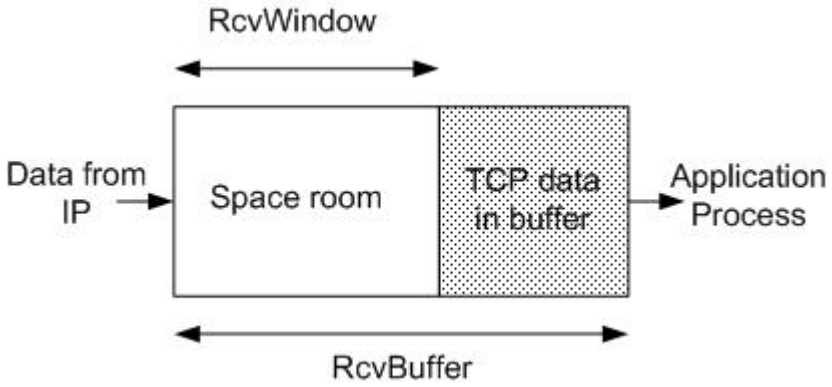


Figure 5. 4: Receiver Buffering

5.4.3 TCP Round Trip Time and Timeout

It is essential to set the TCP timeout value. TCP timeout must be a bit longer than Round Trip Time (RTT). If it is too short, there will be premature timeout and as a result unnecessary retransmissions. On the other hand if it is too long, there will be slow reaction to segment loss. In TCP terminology, SampleRTT is the measured time from segment transmission until ACK receipt (ignoring retransmissions, cumulatively ACKed segments). SamplreRTT varies in order to estimate RTT in a smooth way. To find the proper (adaptive) RTT, several recent measurements are used, not just the current SampleRTT. The general function that provides the EstimatedRTT is the following:

$$\text{EstimatedRTT} = (1\text{-}x) \cdot \text{EstimatedRTT} + x \cdot \text{SampleRTT}$$

This average is an exponential weighted moving average. The influence of a given sample decreases exponentially fast as time passes. A typical value for x is 0.125.

To set the timeout we use EstimatedRTT plus a safety margin. Large variation in EstimatedRTT requires larger safety margin. Timeout is set as follows:

$$\text{Timeout} = \text{EstimatedRTT} + 4 \cdot \text{Deviation}$$

$$\text{Deviation} = (1\text{-}x) \cdot \text{Deviation} + x \cdot | \text{SampleRTT} - \text{EstimatedRTT} |$$

5.4.4 TCP Connection Management

During the request process, TCP sender and receiver establish a connection before exchanging data segments as already mentioned. Sequence numbers and information about buffers and flow control are given. It has been also mentioned that a three way handshaking is used. Here we are going to give some additional details about these steps:

Step 1: Client end system sends TCP SYN control segments to the server and specifies initial sequence number.

Step 2: Server end system receives SYN, replies with SYNACK control segment. Moreover allocates buffer, specifies server and receives initial sequence number.

Step 3: Client also allocates buffers and variables to the connection. Server sends another SYN segment that means that the connection is established.

During the close of a connection four steps are taking place.

Step 1: Client end system sends TCP FIN control segment to server.

Step 2: Server receives FIN, replies with ACK. Closes connection and sends FIN.

Step 3: Client receives FIN, replies with ACK. Enters timeout "wait"

Step 4: Server receives ACK and the connection closed.

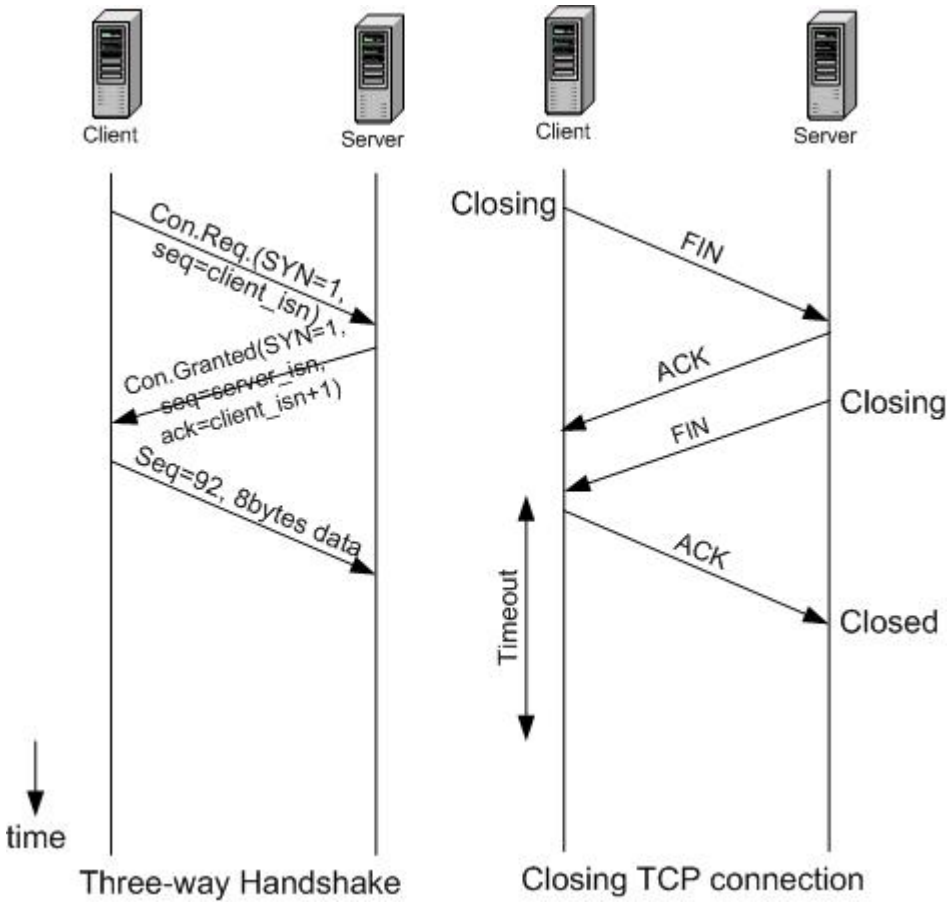


Figure 5.5: Requesting and Closing TCP connection

In the next figures the lifecycle for both client and server are illustrated.

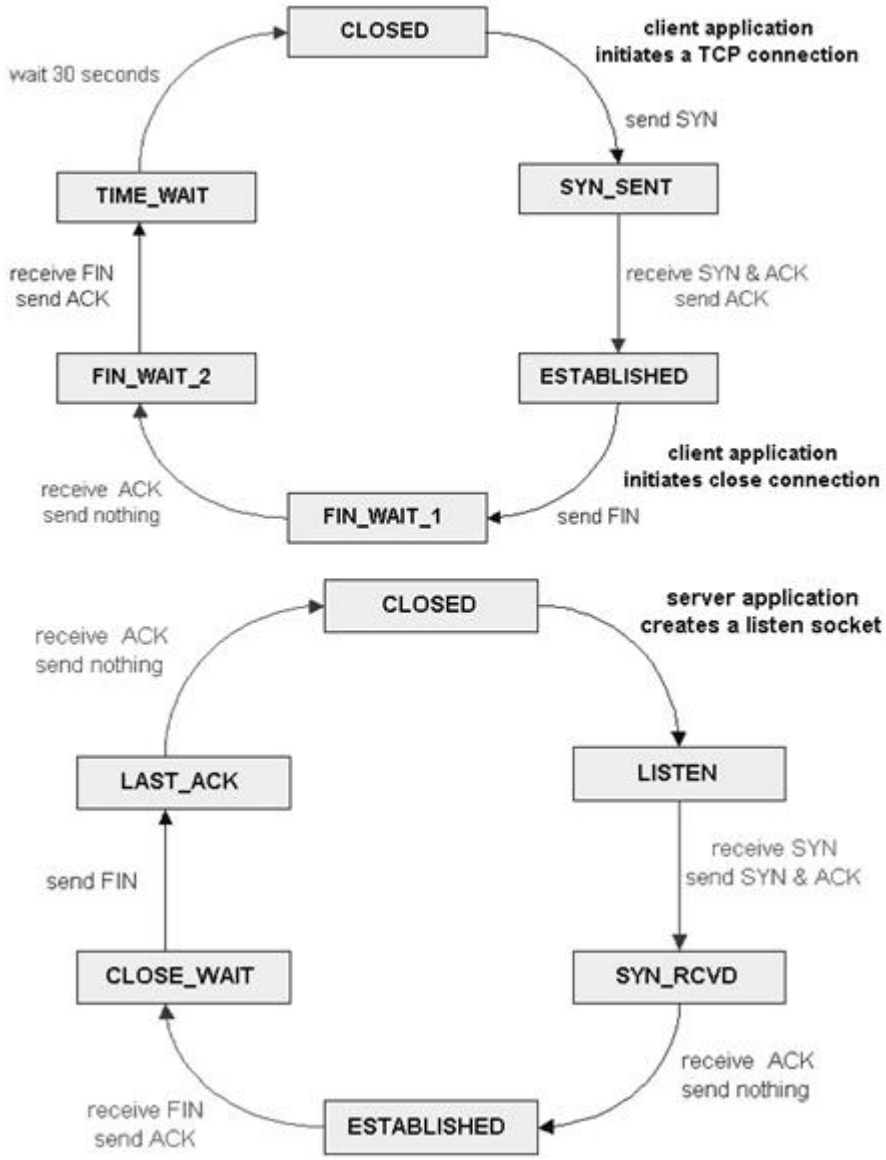


Figure 5.6: TCP client and server lifecycle

5.4.5 TCP Congestion Control

TCP uses end-to-end congestion control. This means that congestion control is not network-assisted. There is no explicit feedback from the network and congestion is inferred from the observed end-systems loss and delay. Using congestion control, transmission rate is limited by the congestion window size (in TCP terminology Congwin) over segments as shown in the following figure:

Figure 5. 7: TCP Congwin

Throughput of w segments, each with MSS bytes sent in one RTT is given as:

$$\text{Throughput} = \frac{w \cdot MSS}{RTT} \text{ bytes/sec}$$

TCP congestion control starts with probing for usable bandwidth. Ideally someone should transmit as fast as possible (Congwin as large as possible) without loss. Instead Congwin increases until loss. When loss occurs, we decrease Congwin and then begin probing (increasing Congwin) again. During this process there are two phases. The first is the slow start and the second is the congestion avoidance. Important variables are Congwin and the threshold that defines the boundary between the slow start phase and the congestion avoidance phase. TCP slow start follows a simple algorithm:

```
Initialize: Congwin=1
For (each segment ACKed)
        Congwin++;
Until (loss event OR Congwin>threshold)
```

Window size is exponentially increased per RTT. Occurrence of a loss event varies and it depends on the TCP version. If we use Tahoe TCP, the loss event is signified by three timeouts. If we use Reno TCP, the loss event is signified by three duplicate ACKs. Tahoe and Reno are the most famous and widely used versions of TCP, however there are many others (adaptive to several environments). Additional details about the various versions we will provide in the next paragraph.

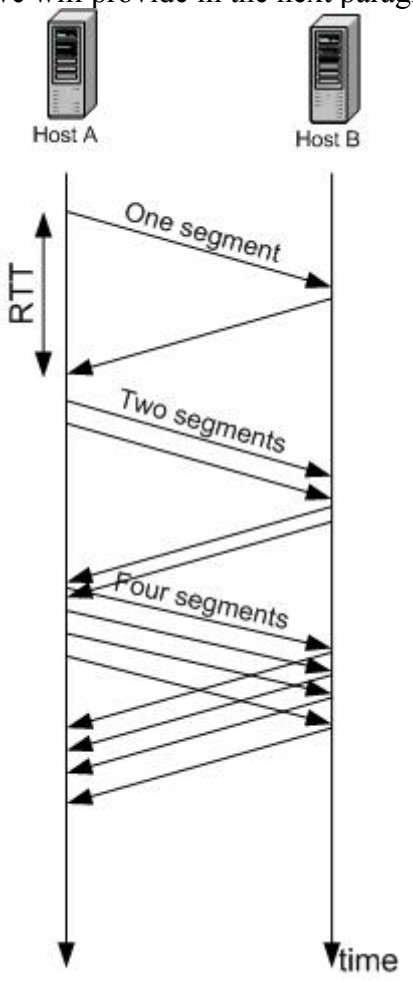


Figure 5.8: TCP Slowstart

Congestion avoidance algorithm is the following:

```
/* slow start is over */
/* Conwin > threshold */

Until (loss event)
                {
                    every w segments ACKed:  Congwin++
                }

threshold = Congwin/2
Congwin = 1
Perform slowstart
```

TCP Reno skips slow start (fast recovery) after three duplicate ACKs. TCP congestion avoidance is also known as Additive increase, multiplicative decrease (AIMD) (after slow start increase window by 1 per RTT and decrease window by a factor of 2 on loss event). Moreover before the threshold, the increase is exponential and after the threshold it is linear as shown on figure 4.9 [12].

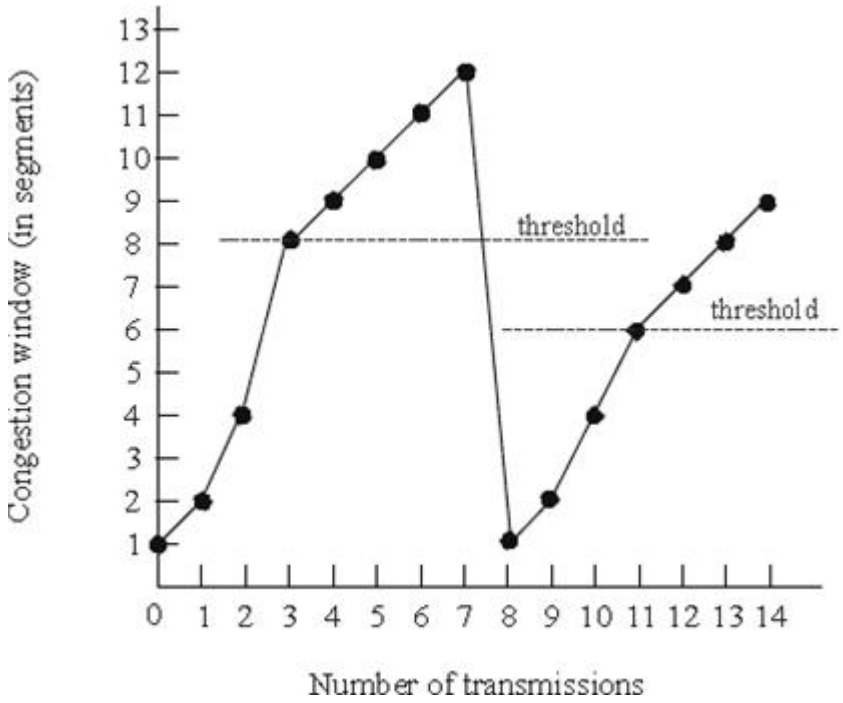


Figure 5.9: Evolution of TCP's congestion window

Using all these tricks, it can be shown that TCP can be characterized as a fair protocol. With fair protocol we assume N TCP sessions sharing the same link (in order to be a fair connection) each should end host must get 1/N of the link capacity. AIMD decreases throughput proportionally as shown in the following figure. That is why TCP can be characterized as a fair protocol.

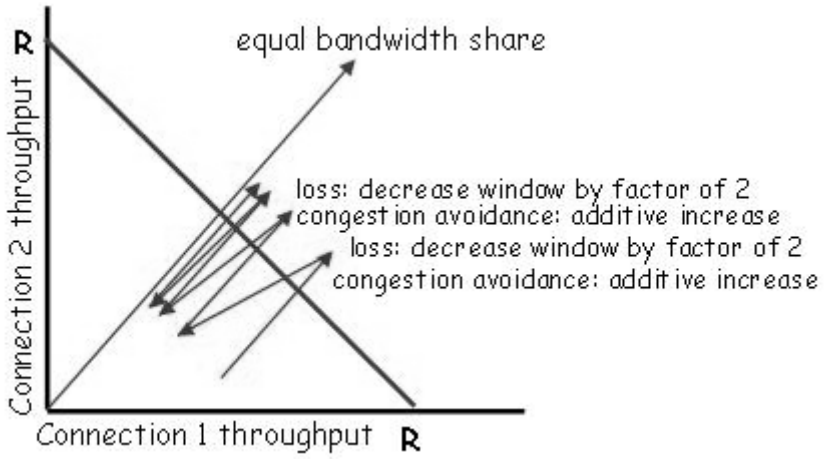


Figure 5.10: TCP Fairness

5.4.6 TCP Tahoe, Reno, Vegas and SACK

The error control mechanism of TCP is primarily oriented towards congestion control. Congestion control can be beneficial to the flow that experiences congestion, since avoiding unnecessary retransmissions can be lead to better throughput-delay tradeoff. TCP utilizes acknowledgments to pace the transmission of segments and interprets timeout events as signs of congestion. In response to congestion, the TCP sender reduces the transmission rate by shirking its window.

There are four major versions of TCP (Tahoe, Reno, NewReno and Vegas). In the following lines we discuss each version in turn [10],[12]:

TCP Tahoe: TCP Tahoe is the oldest version of TCP but on the other hand one of the most famous versions. In the literature is sometimes called just TCP. TCP Tahoe congestion algorithm includes Slow Start and Congestion Avoidance. In order to declare a loss event, three timeouts have to be passed. This is its main drawback [11] as when a segment is lost, the sender side of the application may have to wait a long period of time for the timeout. It implements an RTT-based estimation as was explained earlier.

TCP Reno: TCP Reno (started as a version of TCP Tahoe), except from Slow Start and Congestion Avoidance also includes also Fast Retransmit. In the Fast retransmit mechanism, three duplicate acknowledgements carrying the same sequence number trigger a retransmission without waiting for the associated timeout event to occur. The window adjustment strategy for this early timeout is the same as for the regular timeout and Slow Start is applied. The problem, however, is that the Slow Start is not always efficient, especially if the error was purely transient or random in nature and not persistent. In such case, the shrinkage of the congestion window is, in fact, unnecessary and renders the protocol unable to fully utilize the available bandwidth of the communication channel during the subsequent phase of window re-expansion.

TCP NewReno: TCP NewReno introduces Fast Recovery in conjunction with Fast Retransmit. The idea behind Fast Retransmit is that an ACK is an indication of available channel bandwidth since a segment has been successfully delivered. This, in turn, implies that the congestion window should actually be incremented upon one ACK delivery. Then instead of entering Slow Start, the sender increases its current congestion window by the threshold number. TCP NewReno's Fast Recovery can be effective when there is only one segment drop from a window of data, given the fact that NewReno retransmits at most one dropped segment per RTT. The problem with the mechanism is that is not optimized for multiple packet drops form a single window, and this could negatively impact performance.

TCP Vegas: TCP Vegas approaches the problem of congestion from another perspective. The basic idea is to detect congestion in the routers between source and destination before packet loss occurs and lower the rate linearly when this imminent packet loss is detected. The longer the round-trip times of the packets, the greater the congestion in the routers. Every two round trips delays the following quantity is computed:

$$\rho = (\text{WindowSize}_{\text{Current}} - \text{WindowSize}_{\text{Old}}) \cdot (\text{RTT}_{\text{Current}} - \text{RTT}_{\text{Old}})$$

If $\rho > 0$ the window size is decreased by 1/8 . Else the window size is increased by one minimum segment size. One problem that it does not seem to overcome is the path asymmetry. The sender makes decisions based on the RTT measurements, which, however, might not accurately indicate the congestion level of the forward path. Furthermore, packet drops caused by retransmission deficiencies or fading channels may trigger a Slow Start. However, this problem is common to all of the above versions. Another drawback is that Vega's algorithm is very new (1999) and is not fully embedded in the most popular TCP implementations.

TCP SACK: TCP Selective Acknowledgements (SACK) is a TCP enhancement which allows receivers to specify precisely which segments have been received even in the presence of packet loss. TCP SACK is an Internet Engineering Task Force (IETF) proposed standard which is implemented for most major operating systems. SACK enables receiver to give more information to sender about received packets allowing sender to recover from multiple-packet losses faster and more efficiently. On the contrary TCP Reno and New-Reno can retransmit only one lost packet per round-trip time because they use cumulative acknowledgements.

5.4.7 TCP Latency Modeling

In order to understand the latency of TCP, we have to take into consideration both Static and Dynamic Congestion Window. In Static Congestion Window, the server is not permitted to have more than W unacknowledged outstanding segments. When the server receives a request from the client, immediately sends W segments back-to-back to the client. It then sends one segment for each acknowledgment it receives from the client, until all of the segments of the object have been sent. There are two cases to consider:

1. $W\frac{S}{R} > RTT + \frac{S}{R}$, where

RTT: Round trip time from the server to the client (Excluding transmission time)
S: The MSS in bits
R: Transmission rate of the link from the sender to the client

In this case, the server receives an ACK for the first segment of the first window, before the sender completes the transmission of the window. Therefore, the latency is given by:

Latency = 2 RTT + O/R , where O: Object size in bits

One RTT is required to initiate the TCP connection. Then the client sends a request for the object, and after a total of two RTTs, the client begins to receive data from the server.

2. $W\frac{S}{R} < RTT + \frac{S}{R}$

It can be shown [12] that Latency = 2RTT + O/R + (K-1) [S/R + RTT - $W\frac{S}{R}$] ,

where $K=\left\lceil \frac{O}{S \cdot W} \right\rceil$

Combining we have:

Latency = 2RTT + O/R + (K-1) [S/R + RTT - $W\frac{S}{R}$]$^+$, where $[x]^+ = \max(x,0)$.

In the case of Dynamic Congestion Window, first window contains one segment, the second window two segments, …, the kth window contains 2^{k-1} segments. Let K the number of windows that cover the object, then

$K = \min \{k: 1 + 2^1 + 2^2 + \ldots + 2^{k-1} \geq \frac{O}{S} \} = \min \{ k: 2^k - 1 \geq \frac{O}{S} \}$

$= \min \{ k : k \geq \log_2(\frac{O}{S}+1) \} = \left\lceil \log_2(\frac{O}{S}+1) \right\rceil$

Next we calculate the amount of stall time after transmitting the kth window :

$[S/R + RTT - 2^{k-1} \frac{S}{R}]^+$

therefore the latency is given as:

$$\text{Latency} = 2RTT + \frac{O}{R} + \sum_{k=1}^{K-1} \left[\frac{S}{R} + RTT + 2^{k-1} \cdot \frac{S}{R} \right]^+$$

To simplify the formula, let Q be the number of times the server would stall if the object contained an infinite number of segments :

$$Q = \max \left\{ k : RTT + \frac{S}{R} - \frac{S}{R} 2^{k-1} \geq 0 \right\}$$

$$= \max\left\{k : 2^{k-1} \le 1 + \frac{RTT}{S/R}\right\}$$

$$= \max\left\{k : k \le \log_2\left(1 + \frac{RTT}{S/R}\right) + 1\right\}$$

$$= 1 + \left\lfloor \log_2\left(1 + \frac{RTT}{S/R}\right)\right\rfloor$$

The actual number of times the server stalls is : P = min { Q, K – 1}
Therefore,

$$\text{Latency} = \frac{O}{R} + 2RTT + \sum_{k=1}^{P}\left(RTT + \frac{S}{R} - \frac{S}{R}2^{k-1}\right)$$

$$= 2RTT + \frac{O}{R} + P\left[RTT + \frac{S}{R}\right] - \left(2^P - 1\right)\cdot\frac{S}{R}$$

It is interesting to compare Latency to the minimum latency (the latency that would occur if there were no congestion control, at is if no congestion window constraint exists).

Minimum latency = $2RTT + \frac{O}{R}$

It is easy to show that:

$$\frac{Latency}{MinimumLatency} \le 1 + \frac{P}{\left\lceil \frac{O/R}{RTT} \right\rceil + 2}$$

A few comments are due: if RTT<< O/R, that is if round–trip time is much less than the object's transmission time, then TCP Slow Start will not significantly increase latency. However, with the Web we are often transmitting many small objects over congested end–to-end links, in which case, TCP Slow Start can significantly increase latency. We will have the opportunity to come back to this chapter.

5.5 User Datagram Protocol

User Datagram Protocol (UDP), aside from the multiplexing/demultiplexing function and some light error-checking, it adds nothing to IP. There is no handshaking taking place before sending segments (UDP is connectionless). On the other hand UDP can be comparatively useful (than TCP) due to its special characteristics:

1. No connection establishment. It does not introduce any extra delay to establish a connection. This is why DNS runs over UDP.
2. No connection state in the end systems. A server devoted to a particular application can typically support more active clients when the application runs over UDP rather than TCP.
3. Small packet header overhead. TCP has 20 bytes of header, whereas UDP only has 8 bytes.
4. Unregulated Send Rate. Constrained by the application generating rate, the capabilities of the source, and the access bandwidth to the Internet.

Note however, that the receive rate can be limited by the network congestion even if the sending rate is not constrained. Furthermore, the lack of congestion control in UDP is a serious problem when running streaming multimedia applications over it. If everyone were to start streaming high-bit rate video without using any congestion control, there would be so much packet overflow in the routers that no one would see anything. Researchers have proposed new mechanisms to force all sources, including UDP sources, to perform adaptive congestion control. Many of today's proprietary streaming applications, run over UDP, but they have built acknowledgement and retransmissions into the application in order to reduce packet loss. That is, the application process can communicate reliably without having to succumb to the transmission rate constraints imposed by TCP's congestion control mechanisms. TCP is used as an out of bound service and transports control signals reliably.

The 8-byte UDP segment structure consists of Source and a Destination port numbers: 4 bytes long and a 2-byte checksum (used by the receiving host to check if errors have been introduced into the segment). UDP at the sender performs the one's complement of the sum of all 16-bit words in the segment. The result is put in the checksum field of the UDP segment. At the receiver, all 16-bit words of the segment are added, including the checksum. If no errors are introduced into the segment, the sum at the receiver will be all 1's. The reason why UDP provides a checksum in the first place is that there is no guarantee that all the links between source and destination provide error checking. IP is supposed to be able to run over just about any layer-2 protocol (that is why is used as a peer communication transfer protocol between routers) thus UDP does nothing to recover from an error.

5.6 Real Time Protocol

Real Time Protocol (RTP) specifies a packet structure for packets carrying audio and video. RTP provides payload type identification, packet sequence numbering and timestamping. It runs in the end systems and runs on top of UDP. RTP does not provide any mechanism to ensure timely delivery of data or provide other quality of service guarantees because RTP encapsulation is only seen at the end systems and not by intermediate routers. In assistance with RTP, Real Time Control Protocol (RTCP) is used. Each participant in an RTP session periodically transmits RTCP control packets to all other participants. Each RTCP packet contains sender and/or receiver reports that reports statistics useful to the application. RTCP is also used for the synchronization of Streams and for bandwidth reservation. Typically

RTCP traffic is 5% of RTP traffic, from which 75% is reserved for the receiver and 25% for the sender. At last but not least RTCP runs over TCP.

5.7 Discussion: Unwiring the Internet

Even today's wired internet has so many problems that some people persist to call it the World Wide Wait. These problems will enlarge in the future Mobile Internet. Although TCP was introduced years ago on the wired internet it will continue to be THE transfer protocol. Its hard defensive behavior towards any kind of loss will be its weakest point over wireless and mobile networks. TCP can not infer if an error is due to traffic congestion or losses over a wireless link. Our approach to this problem is to find the proper version of TCP to each profile of application and user over the UMTS network. Notice that the UMTS network is implemented using TCP and end-to-end QoS.

CHAPTER 6: Simulation Parameters & Simulator

6.0 Introduction

The UMTS network is not just another wireless telecommunication system. Its aim is to converge wireless/mobile networks and internet services in a single network. UMTS users will be mobile and some of them fast mobile, that is why de facto solutions for wired internet and typical wireless infrastructures will just not be enough. In this chapter we will explain how we developed our simulation platform in order to simulate a realistic UMTS environment and the behavior of various versions of the Transmission Control Protocol which are standards from the Internet Engineering Task Force over UMTS.

6.1 Network Parameters

In the following table we illustrate the major parameters of the UMTS R4 network , in a city environment ,which will be "on air" in the following year according to the 3GPP which is an international organization for the standardization of UMTS [5],[6].

PARAMETER	VALUE
Number of cells	7 (hexagonal)
Cell Side	200m
Path loss propagation exponent β	3.5
Path loss propagation parameter A	–30dB
Shadowing parameter σ	4dB
Number of oscillators (Jakes)	8
Number of multipath rays	5
Noise	-132 dbW
Doppler frequency	0,4,20,80 Hz
Chip rate	3.84 Mcps
PC frequency	1500 Hz
PC power range	80dB
PC step	0.5 dB
PC SIR objective	2.5 to 3 dB
PC loop delay	1,2,3,4,10
Maximum TX Power	-16 dBW
Packet length(MTU)	1000 bytes

Table 6. 1: UMTS Simulation Parameters

We examine the downlink behavior because internet services are highly asymmetric, but similar results are taken if we examine uplink behavior as well [5]. We also use a downlink data rate per user of 120Kbps and 240Kbps. The reason we made this choice is because, although UMTS network provides downlink data rate up to 2084kbps, operators will establish data rate limits per user in order to avoid high traffic conditions. For power control we use the UMTS forum choice which is close loop power control.

In order to introduce users in the network, at first we have to calculate the cell capacity.

The spread factor, as described in section 3.4 chapter 3, for Admission Control that attaches 120Kbps per user is :

$$G_P = \frac{3840000}{120000} = 32 = \text{Spreading Factor}$$

And hence for E_b/N_o = 3dB (2.5+0.5dB) noise

$$X \approx \frac{G_P}{E_b / No} = \frac{32}{3dB} \approx \frac{32}{2} = 16 \text{ users per cell (with one TRX per cell)}$$

For a 7 cell network the number of users will be approximately equal to 110. If we design a network with 3 TRXs/cell the total number of users that this architecture can handle is approximately equal to 330, but the performance of services and especially TCP will be poor if there will be over 110 active users and very poor if there will be over 160 active users [6]. By active users we mean clients that use internet services simultaneously with others in a shared environment.

Moreover we have to determine the RTT not only between RNC and UE but also between RNC and SGSN. The average RTT between SGSN and RNC is 20 msecs and the average RTT between SGSN and the server is 80 msecs. Moreover the approximate RTT between RNC/BS and SGSN is 10 msec [22].The approximate RTT between BS and UE is equal to

$$RTT = TRANS_P + TRANS_A + 2 \cdot PROP + 2 \cdot PROC$$

$$\approx \frac{1000 \cdot 8}{120000} + \frac{40 \cdot 8}{120000} + 2 \cdot \frac{400}{3 \cdot 10^8}$$

= 100 msec

In order to provide a starting point for the TCP to adapt we set RTT=120 msec at the begining of the simulation

Following the same methodology with the one proposed above, we can calculate cell capacity for users with 240Kbps downlink data rate.

In such case we obtain:

$X \approx \frac{G_P}{E_b / No} = \frac{16}{3dB} \approx \frac{16}{2}$ =8 users per cell (with one TRX per cell)

For a 7 cell network the number of users will be equal to approximately 55. If we design a network with 3 TRXs/cell the total number of users that this architecture can handle is approximately 160 but the performance of services and especially TCP will be poor if there will be over 60 active users and very poor if there will be over 100 active users [6]. Moreover RTT = $TRANS_P$ + $TRANS_A$ + 2 · PROP + 2 · PROC = 50 msec and we mitially set RTT = 80msecs.

6.2 Error Statistics over UMTS (WCDMA) air interface

While trying to simulate UMTS, the main objective is a characterization of the error process. Lets consider data blocks (for a constant packet length), where a block is the data unit handled and can made up of one or more radio frames. Let X(i) = E if the block I is in error (which occurs with probability $P_e(i)$) and C if it is correct. For an ergodic process X(i) we have [6]:

P[burst length ≥ k] = P[X(t)=E,t=2,…,k|X(0)=C,X(1)=E]

$$= \frac{P[X(0) = C, X(t) = E, t = 1,..,k]}{P[X(0) = C, X(1) = E]}$$

$$= \frac{\beta(k)}{\beta(1)}, \text{k} \geq 1$$

where $\beta(k) \approx \frac{1}{N-k} \sum_{i=1}^{N-k} (1 - P_e(i-1)) \prod_{j=1}^{k} P_e(i+j-1)$

From the traces of the error probabilities as obtained by simulation we can estimate various statistics. The above model can be described by a Two-State Markov error model as shown in the following Figure [6].

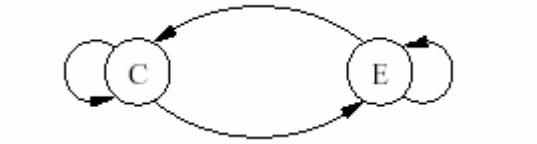


Figure 6. 1: Two-state Markov error model

The model is fully characterized by its transition matrix:

$$P = \begin{bmatrix} p_{CC} & p_{CE} \\ p_{EC} & p_{EE} \end{bmatrix}$$

Where p_{CE} is the transition probability from correct to erroneous i.e. the conditional probability that a packet is in error given that the previously transmitted packet was correct, and all other entries are similarly defined. The balance equation is:

$P(E) \cdot p_{EC} = P(C) \cdot p_{CE}$

and after replacing $P(C) = 1 - P(E)$

we obtain

$P(E) = \frac{p_{CE}}{p_{CE} + P_{EC}}$, which is the average error rate

Moreover $P(E|E) = p_{EE} = 1 - p_{EC}$

Furthermore the probability of non-isolated errors is given in the following:

P[burst length>1] = P[E|E] = p_{EE}
and conditional probability of remaining in bad state is given by

$P(E|E) = p_{EE} = \frac{p[burstlength > k]}{p[burstlength > k - 1]}$, k>1

Note that the latter two quantities are equal in this case. As a special case for independent errors we have P[E]=P[E|E].

Simple as it is the two-state Markov Model does not capture all types of behavior. An obvious way to get around this problem is use a multistate Markov Model. However unlike the case of two states where average error rate and burstiness uniquely identify the model, in this case there may be multiple models fitting a set of burstiness parameters since a N-state Markov chain has N·(N-1) independent parameters in general. We therefore further restrict ourselves to a three-state model such as the one shown in the next figure [6].

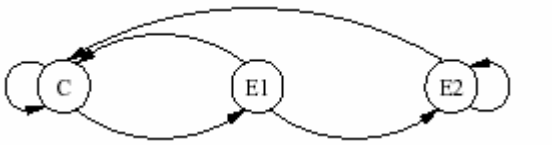


Figure 6. 2: Tree-state Markov error Model

Two error states are now present and the second of them, E2, can only be entered from the first, E1. Also exiting E2 necessarily, leads to the Correct State (not to E1). The transmission matrix for such a model is as follows:

$$P = \begin{bmatrix} p_{CC} & p_{C1} & 0 \\ p_{1C} & 0 & p_{12} \\ p_{2C} & 0 & p_{22} \end{bmatrix}$$

where the zeros correspond to the disallowed transitions.
The value of the parameters already considered for the following two-state models are in this case:

The average error rate is

$P[E]=P[E1]+P[E2]=a_0$,

where $P[\text{burst length}>1]=p_{12}= a_1$,

$$\frac{p[burstlength > 2]}{p[burstlength > 1]} = p_{22} = a_2$$

We know that, for independent errors $a_0=a_1=a_2$

For the two-state Markov model considered earlier $a_1=a_2 \neq a_0$

For the three-state Markov model, we do not have any clear relationship among a_0,a_1,a_2 .

Moreover following the balance equation analysis we find that

$P(C)=P(E1)/p_{C1}$

where $p_{C1}=\frac{P(E)\cdot(1-p_{22})}{(1-P(E))\cdot(1-p_{22}+p_{12})}$,

$$P(E1)= \frac{P(E)\cdot(1-p_{22})}{(1-p_{22}+p_{12})}$$

and P(E2)=1-P(C)-P(E1)

The following Figures are from [6]. They show the probabilities that the length of an error burst exceeds k with $1 \leq k \leq 4$, along with the average error rate. In all graphs the x-axis denotes the user ID numbers are and users are ordered according to decreasing values of P[burst>1]. Three cases of fading rate are shown namely, f_D=6 , 20 and 80 Hz.The case of static fading corresponds to independent errors since after some transient everything reaches steady-independent and the SIR seen by each user is constant.

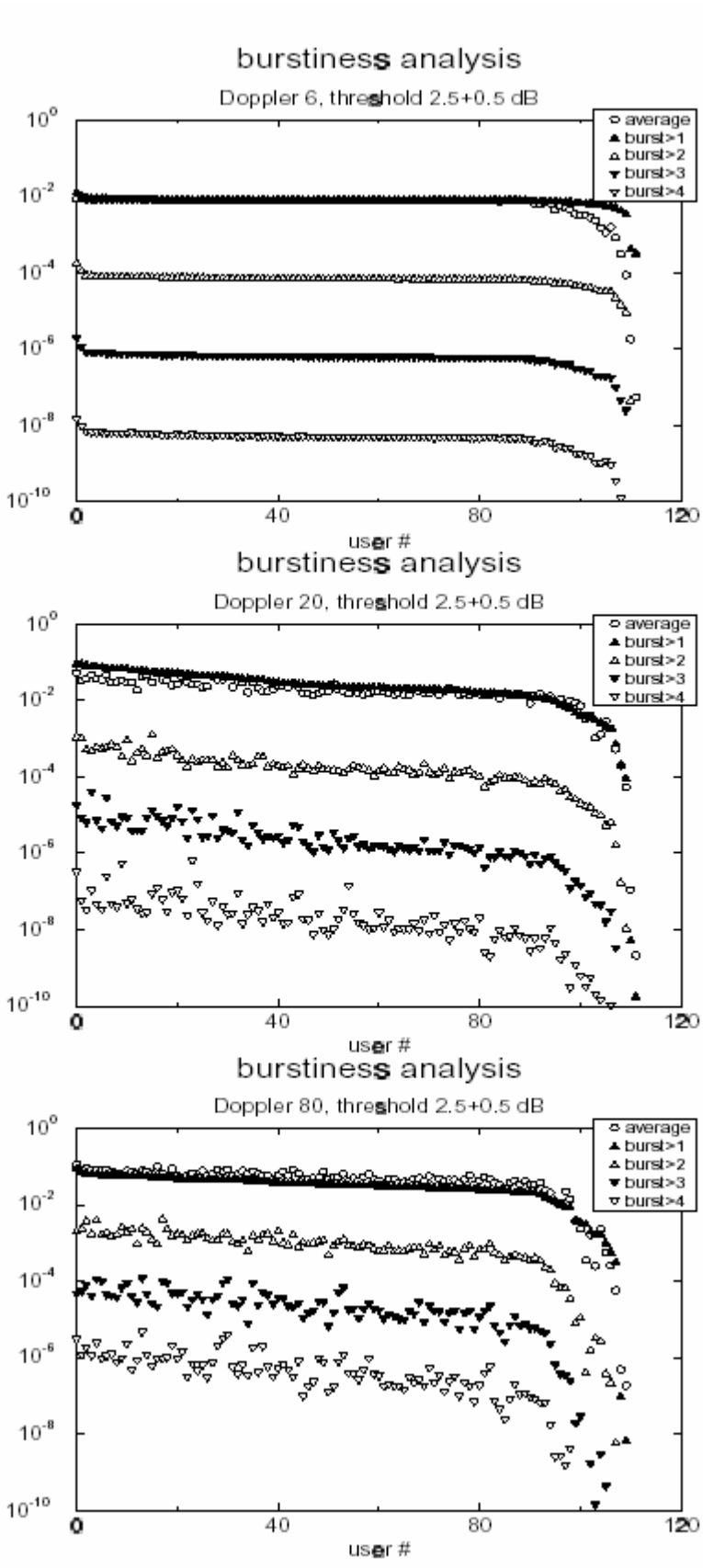


Figure 6. 3: Error burst statistics

The results in the first plot, for slow fading, show that most users will see independent errors, since power control equalizes SIR so well that all randomness due to fading and interference is absorbed. Also, the fading randomness equalizes performance across most users, since all will essentially see the same average. For them error rate and transition probability are not equal, whereas on the other hand it seems that higher order burst probabilities decrease by a fixed amount, which resembles the behavior observed in the two-state Markov model.

As expected, as fading rate increases everything gets more mixed, but a similar qualitative trend can still be observed. Note that a positive burst correlation can be observed i.e. P[E|E] ≈ P[burst>1]>P[E] in most cases for values of the Doppler up to 20Hz (second plot). This indicates that the system tends to stay in the bad state i.e. errors tend to be clustered, which is intuitive pleasing since the channel has memory. This resembles to the behavior observed in the three-state Markov model.

On the other hand, for fast fading (see the last plot) the opposite is observed, i.e. the system tends to escape from error states. This can be explained by noting that the power control tends to increase the transmitted power (related work [5] has shown, that the optimal handover is observed at 2.5 to 3.5 dB) when an error is observed while at the same time the dynamics of the channel tend to make the channel exit quickly from bad conditions, so that right after an error the probability that a transmission is successful is higher than average. We refer to this case in which errors tend to occur isolated as the case of "anticorrelated" errors. This resemble the behavior observed in the two-state Markov model.

The above behavior can be also checked by simple calculations a_0,a_1 and a_2 for each case based on the formulas in this section. In the following Figure the P[E|E] results corresponding to the simulation cases shown before are presented.

The first plot shows that for slow fading most users experience close-to-nominal performance and that for the others a well-defined line can be identified. For faster fading (see the next plot) more dispersion is observed, as expected but there is still a definite trend as to where points are located. Furthermore most points indicate error correlation, as already observed when illustrating burstiness results. Interestingly, as fading rate is further increased to 80dB, the dispersion is reduced, since most differences are averaged out, so that the "cloud" of points becomes more compact.

To sum up, the two-state Markov model best describes slow users like these with doppler frequency equal to 6Hz, as well as very fast users (after power control has been activated) with Doppler frequency equal to 80Hz. On the other hand three-state Markov model best describes users that are not very slow or very fast, with Doppler frequency equal to 20Hz.

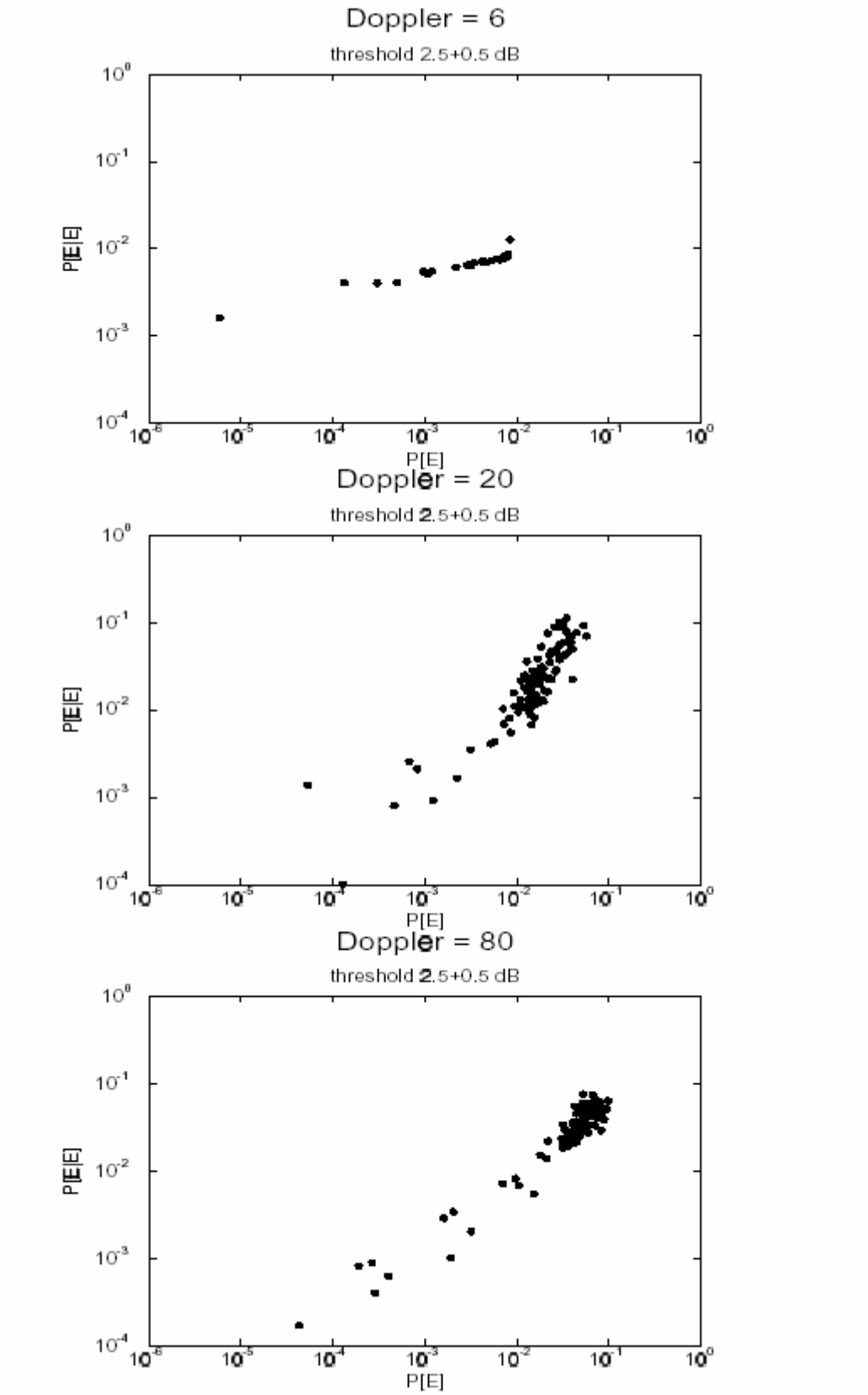


Figure 6. 4: P[E|E] vs. P[E]

6.3 TCP improvements proposed for 3G

3GPP, research groups [14] as well as pioneer companies in the area of 3G mobile networks [22] have studied TCP over UMTS and proposed the following improvements which must be taken into account in any simulation:

1. Large window size at both Sender and Receiver (This must be done in order to eliminate the bad impact of high error rate in wireless links).

2. Increased Initial Window for sender: Senders can avoid delayed ACK mechanism (timeouts) of TCP by using a larger initial window. It can be shown [23] that the initial CWND must be equal to:

   min (4 MSS , max (2 MSS , 4380 bytes))

3. Limited Transmit at Sender : extends Fast Retransmit/Fast Recovery for TCP connections with small congestion windows that are not likely to generate the three duplicate acknowledgements required to trigger Fast Retransmit. If a sender has previously unsent data queued for transmission, the limited transmit mechanism calls for sending a new data segment in response to each of the first two duplicate acknowledgments that arrive at the sender. This mechanism is effective when the congestion window size is small or if a large number of segments in a window are lost. This may avoid some retransmissions due to TCP timeouts.

4. IP Larger than Default: The maximum size of an IP datagram supported by a link layer is the MTU (Maximum Transfer Unit). The link layer may, in turn, fragment IP datagrams into PDUs. For example, on links with high error rates, a smaller link PDU size increases the chance of successful transmission. With layer two ARQ and transparent link layer fragmentation, the network layer can enjoy a larger MTU even in a relatively high BER (Bit Error Rate) condition. Without these features in the link, a smaller MTU is suggested. TCP over 2.5G/3G should allow freedom for designers to choose MTU values ranging from small values (such as 576 bytes) to a large value that is supported by the type of link in use (such as 1500 bytes for IP packets on Ethernet). Given that the window is counted in units of segments, a larger MTU allows TCP to increase the congestion window faster. Typical value is 1000 bytes (also provides small overhead 12bytes of head per 1000 bytes total equal to 1,2%).

5 Path MTU Discovery at Sender and Intermediate Routers: allows a sender to determine the maximum end-to-end transmission unit.

6. Explicit Congestion Notification enabled at Sender, Receiver & Intermediate Routers: Explicit Congestion Notification, RFC3168, allows a TCP receiver to inform the sender of congestion in the network by setting the ECN-Echo flag upon receiving an IP packet marked with the CE bit(s). The TCP sender will then reduce its congestion window. Thus, the use of ECN is believed to provide performance benefits.

6.4 Simulator

In order not to "reinvent the wheel", as base for our simulations we use the widely used Network Simulator (NS) version 2 (2.1b9). NS2 is an object-oriented, discrete event driven network simulator developed at Lawrence Laboratories - UC Berkely (maintained by Information System Institute-ISI) written in C++ and OTcl and running in the UNIX operating system. It implements network protocols such as TCP and UDP, traffic source behavior such as FTP, Telnet, Web and router queue management mechanism such as Drop Tail. A simplified User's View of NS is the following [25]:

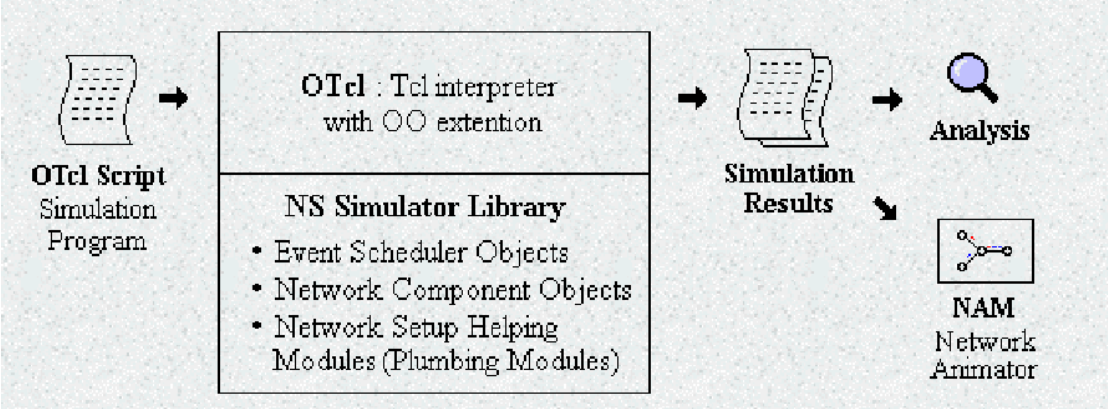


Figure 6. 5: A simplified User's View of NS

As shown in the above figure all simulation results can be viewed by the Network Animator (NAM) which is a simulation display tool.

In order to create a simulation topology we have to create simulation nodes. Then we have to determine which nodes are routers, which of them are servers (sources) and which are clients. After that, we determine how these nodes are connected to each other and RTT is estimated as well as bitrate of each connection. Then we determine the logic connection (agents) between servers and clients and the protocol they use to communicate (TCP/UDP). Last, we attach a service on each server and a "sink" on each client. This services inherit a traffic generator (FTP, HTTP, Telnet). In our simulations when the simulation starts all users are uniformly distributed in all cells and the request arrival process is Poisson. We chose to stop each run after 600seconds of operation. Due to the timestamp-oriented nature of NS, when a new simulation run begins the user of the simulator is asked to provide a random number.

At the beginning of each simulation a different two-State Markov scenario is chosen for each network user according to figure 6.3 for Doppler frequency of 6 and 60 Hz and a three-state Markov scenario for Doppler Frequency of 20 Hz. Moreover the user of the simulator has to determine at the beginning of each run the number of active network users and the type of TCP that must be simulated (TCP(Tahoe), TCP/Reno , TCP/New-Reno, TCP/Vegas, TCP/Sack1). We do not investigate the behavior of split protocols (e.g. Snoop [19]) because these protocols require modifications at the base station and once they are developed it is practically difficult to change or upgrade them. That is why those protocols are used mainly in fixed wireless networks like Wireless Local Area Networks [13]. Moreover end-to-end solutions solve the problem how to design and implement integrated networks to work

with a variety of heterogeneous networks (mixed wired and wireless) and devices. Furthermore research projects [26] have shown than slit solutions radio to radio resource waste. Finally we have to investigate end-to-end behavior because as shown in Figure 4.1, end-to-end QoS is used (for transport protocols as well).

All the results (the output) of the simulator are saved in a standard file (out.tr.gz). When the simulation finishes we unzip that file and using scripts written in cat and awk (two pattern recognition tools of UNIX) we find the number of Acked packets and the total number of Send Packets. In order to check the behavior of our simulator we can create a file (out.nam) which is input for NAM and view the simulator's behavior step-by-step.

Our simulator can be summarized in the following Figure (where also a snapshot of the simulator for 100 users, F_D=20 Hz using TCP/NewReno and Telnet, is shown).

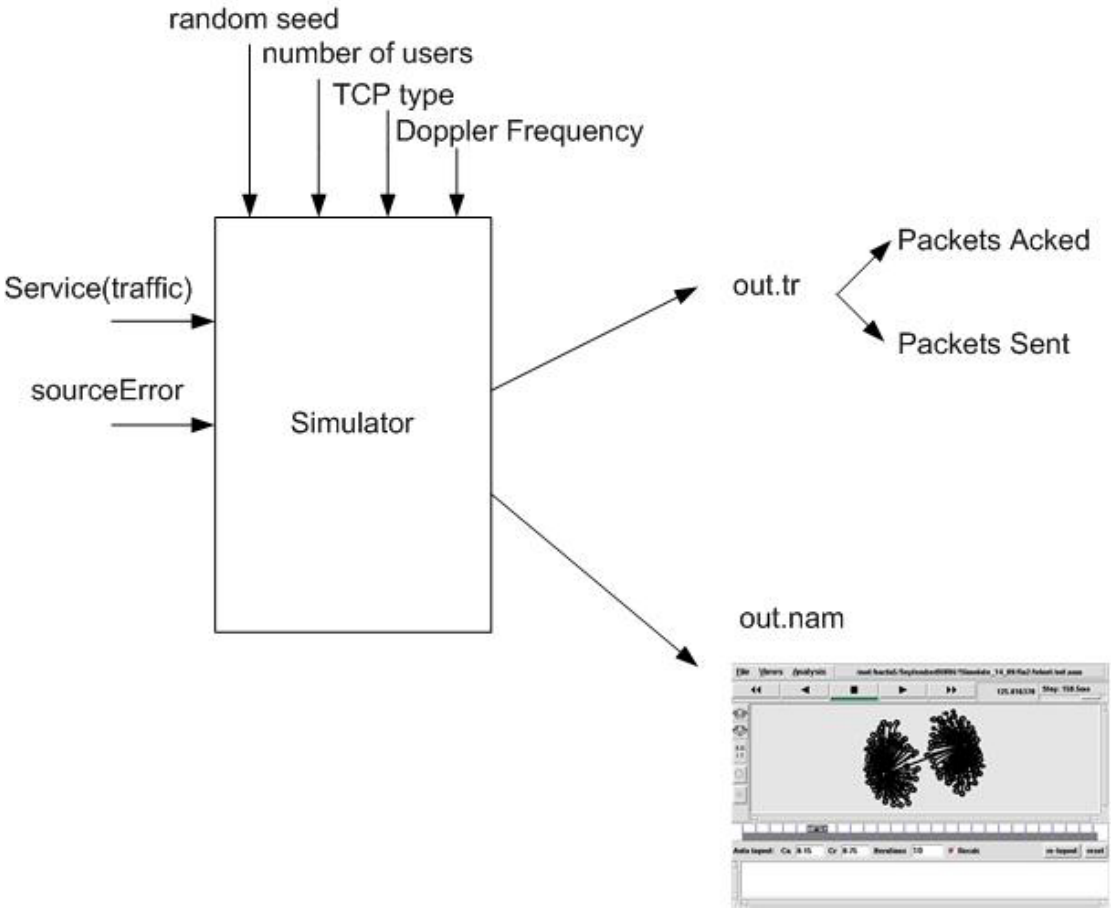


Figure 6. 6: Simulator Overview

6.5 Traffic Specifications

We investigate the behavior of three well known and widely used application protocols: File Transfer Protocol (FTP), Hyper Text Transport Protocol (HTTP) and Telnet. These applications run over TCP, because they require reliable transfer of

information, and are characterized as On-Off applications. The most demanding in terms of volume is the FTP and the next two follow.

Both FTP and Telnet starts from a specific server on one edge of the Internet and turn out to a client on the other edge. FTP transmits files. Only congestion and cancellation by the user affect it. It normally finishes when the entire file is transferred. As a result FTP is highly asymmetric. On the other hand Telnet is symmetric. Its aim is to transfer commands from one edge to another. When the command is received at the destination it must also be received back at the source. NS2 contains a special traffic empirical generator called Application/Telnet which simulates the behavior of Telnet after making a special profile for each user (client node) attached in the simulated network.

The most complex of the above protocols is HTTP (HTTP/1.1 clients opens only one socket with the server). NS2 approach for simulating HTTP is different than that for FTP and Telnet. A number of nodes are reserved as servers. These servers comprise the server pool. On the other hand clients demand pages from the server pool. Moreover a proxy server stands between clients and server pool. The traffic model we used is the http-mod.tcl which is the defacto model for http simulations in NS. This model is adapted according to the number of users and the simulated environment and is attached to each user. We used the wireless format of this model which has the following characteristics: Inter-page interval is exponentially distributed with mean duration 1second, Number of Objects per Page is equal to 4. Inter object interval exponentially distributed with mean duration 0.01 seconds, and the number of packets per object is Pareto distributed with shape parameter equal to 1.2 and scale equal to 10.

For UMTS routers we used the DropTail queue. Droptail inherits round robin functionality and keeps a buffer size equal to 110. When the buffer is full the queue "drops" the arriving packet.

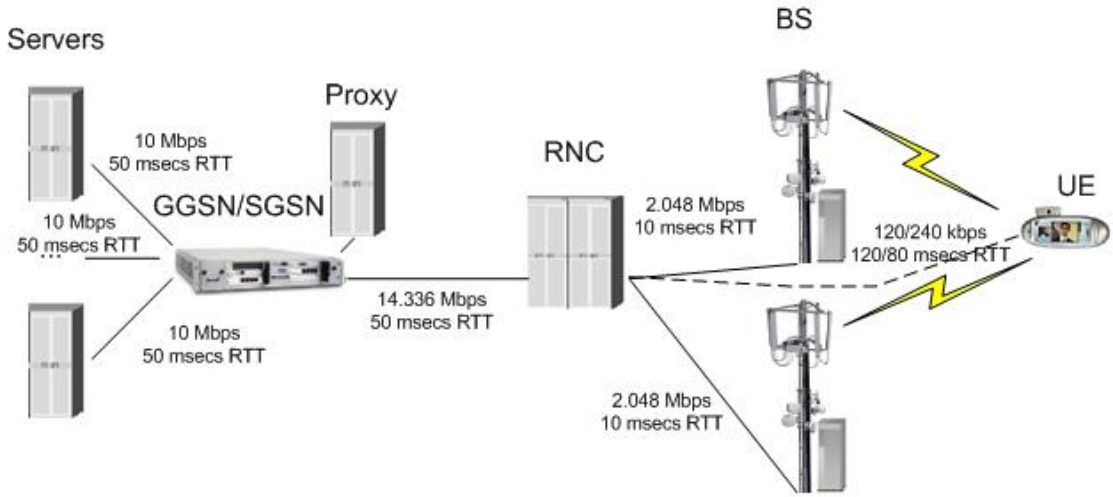


Figure 6. 7: Simulation Topology

CHAPTER 7: Results & Discussion

7.0 Introduction

In this chapter we present and discuss the results from our simulator. For each application we show two types of graphs for 120Kbps reserved bitrate as we will explain shortly. We refer to the corresponding section as the TCP Performance section. Furthermore we investigate the efficiency of all types of TCP in a application-oriented approach. We refer to the latter section as the TCP Efficiency section.

7.1 TCP Performance

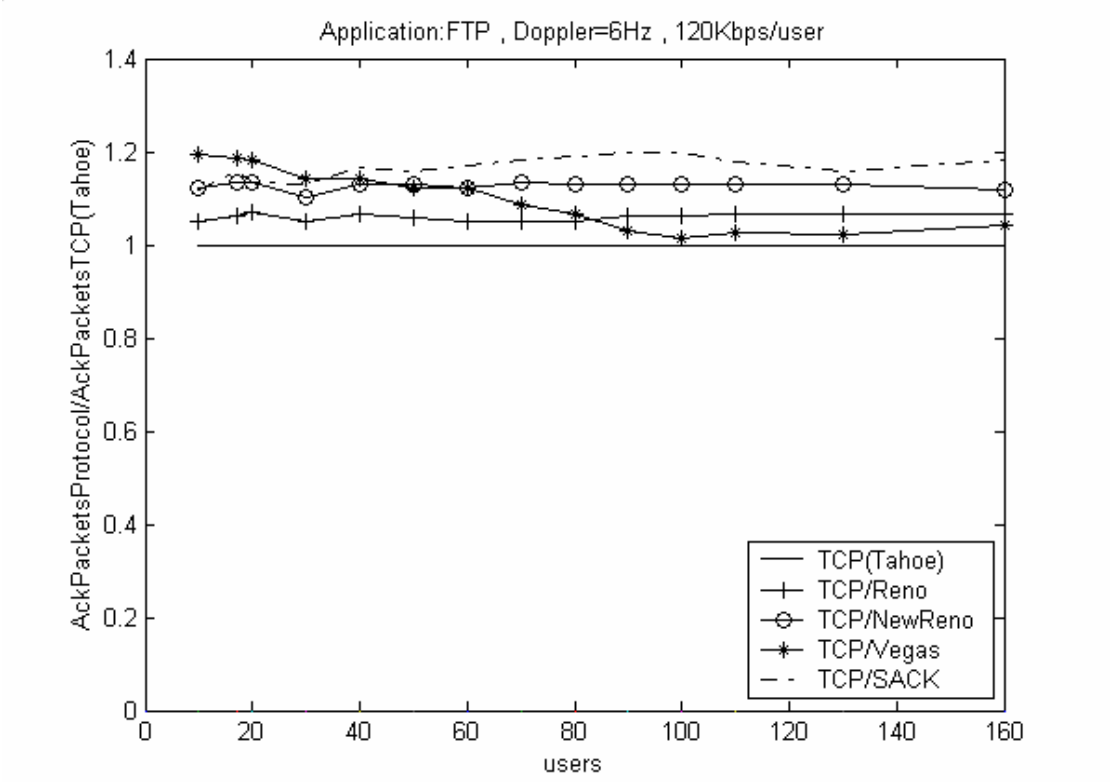
We investigate the behavior of TCP Tahoe, TCP Reno, TCP New-Reno, TCP Vegas and TCP SACK over UMTS for the following services: FTP, Telnet and HTTP for both 120Kbps and 240Kbps admitted users. In this order we illustrate the results of our simulations in the following graphs.

We use two types of graphs. The first illustrates the Total number of Acked Packets form the receiver to the sender of each type of TCP divided by the total Total number of Acked Packets from the receiver to the sender of TCP Tahoe. By Total number of Acked Packets we mean those packets whose Ack has been received by the sender. We clarify this because some Acks may be lost due to traffic congestion or errors over lossy wireless links. The X-axe denotes the total number of users in each simulated case.

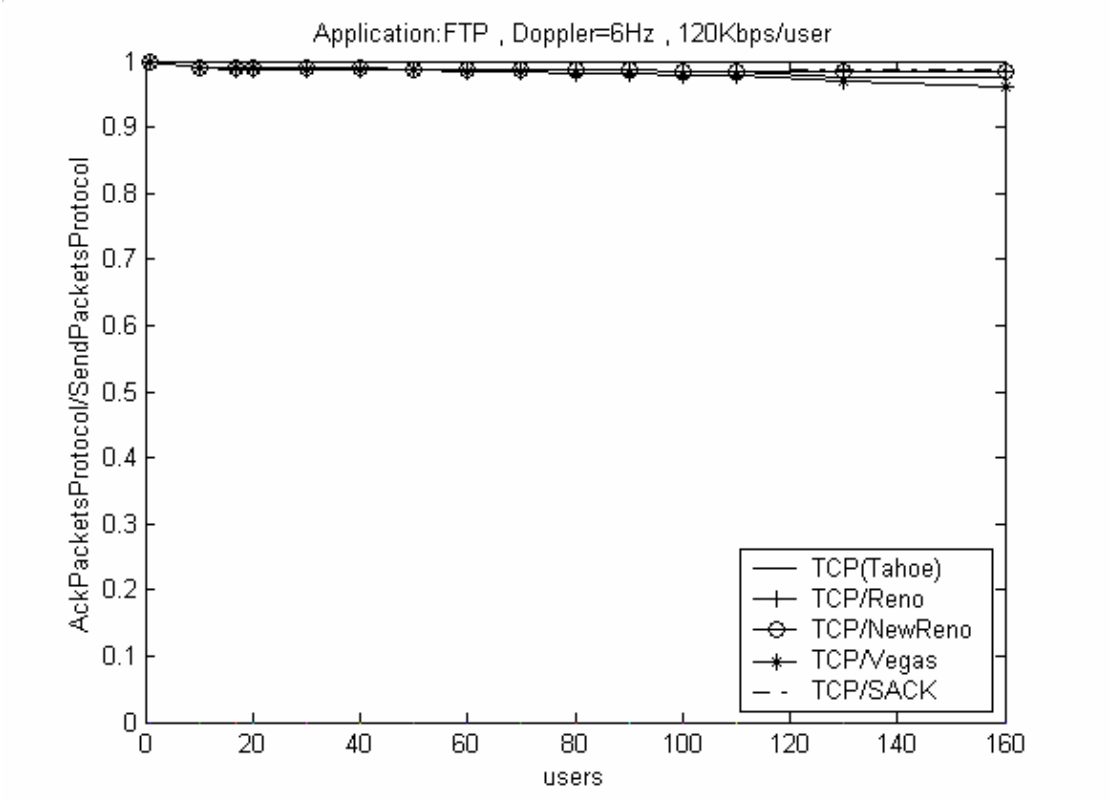
The second type of graphs give us an indication of Energy Consumption (as well as Packet sending efficiency). They show the fraction of the Total number of Acked Packets from the receiver to the sender for each type of TCP divided by the Total number of Packets sent from receiver to the sender for each type of TCP. By Total number of Packers sent includes those packets that never reached the client due to traffic congestion or transmission errors. X-axis denotes the total number of users in each simulated case.

We decided to run each simulation five times with a different seed each time and we report the average value of these runs in our graphs. In order to be compatible with the error statistics section of the previous chapter, for 120Kbps users the total number of users is between 10 and 160 and in the case of 240Kbps users the total number of users varies between 10 and 100. In the sequel we discuss the behavior of each version of TCP and we propose the optimal version or versions for the UMTS environment. For the construction of the graphs we used the MATLAB tool.

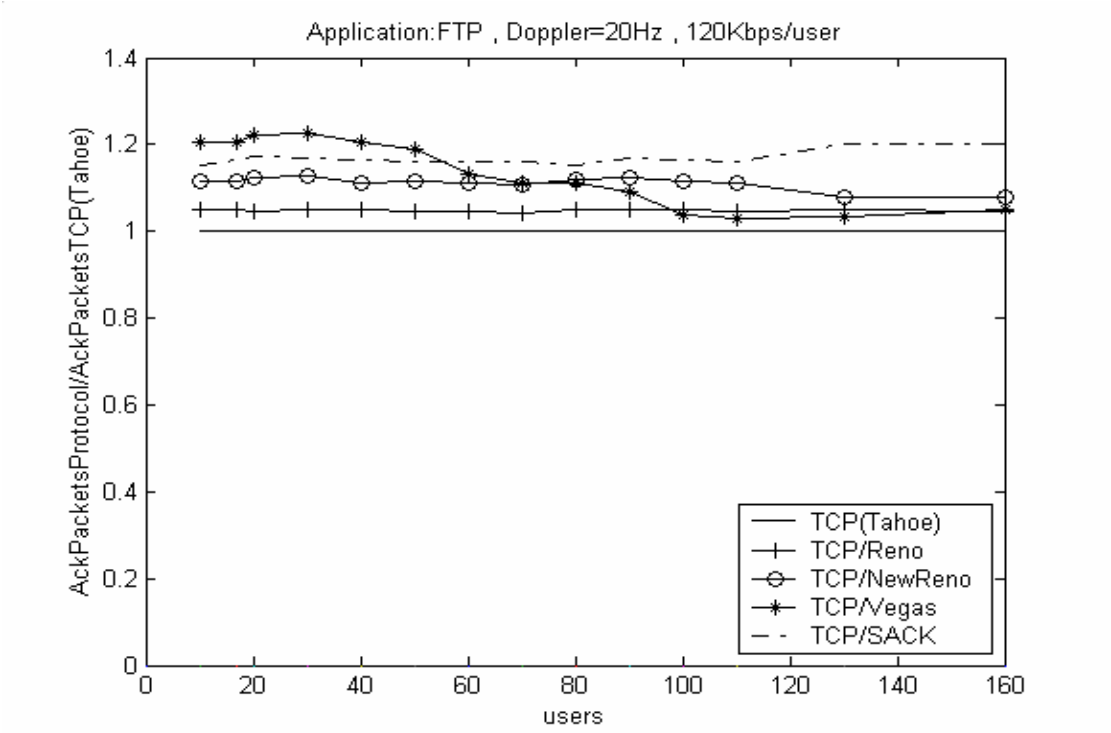
7.1.1 FTP Application



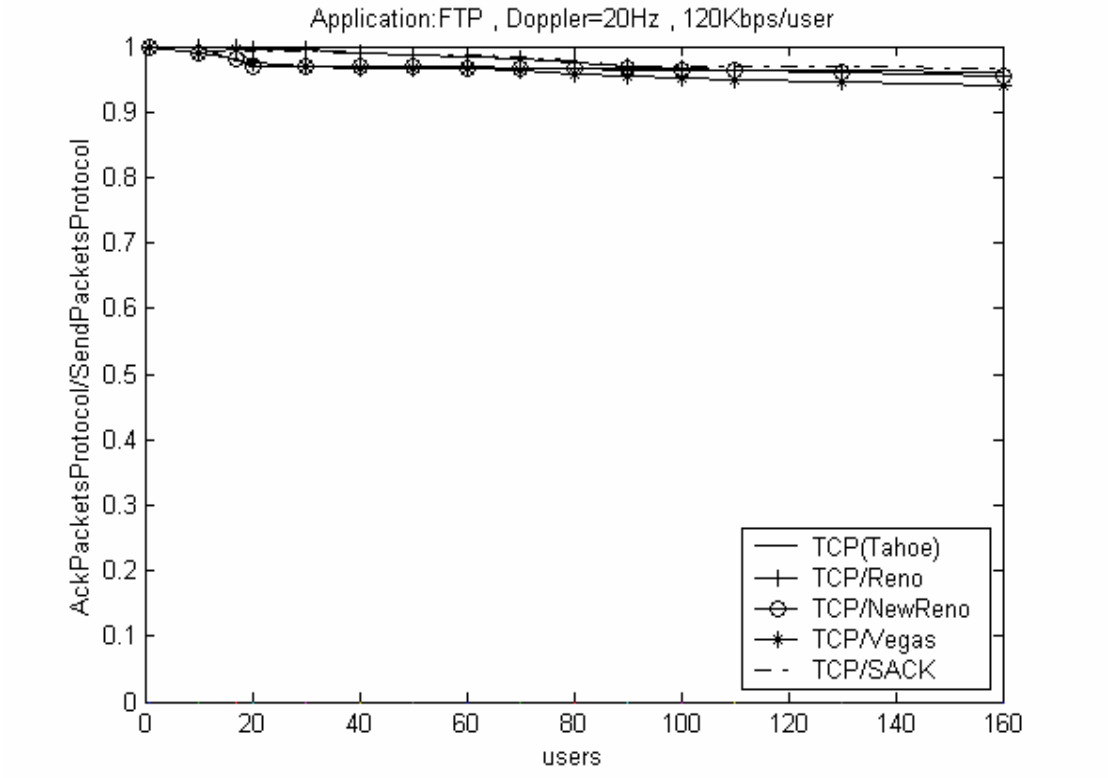
Graph 7. 1: Performance of FTP/6Hz/120Kbps



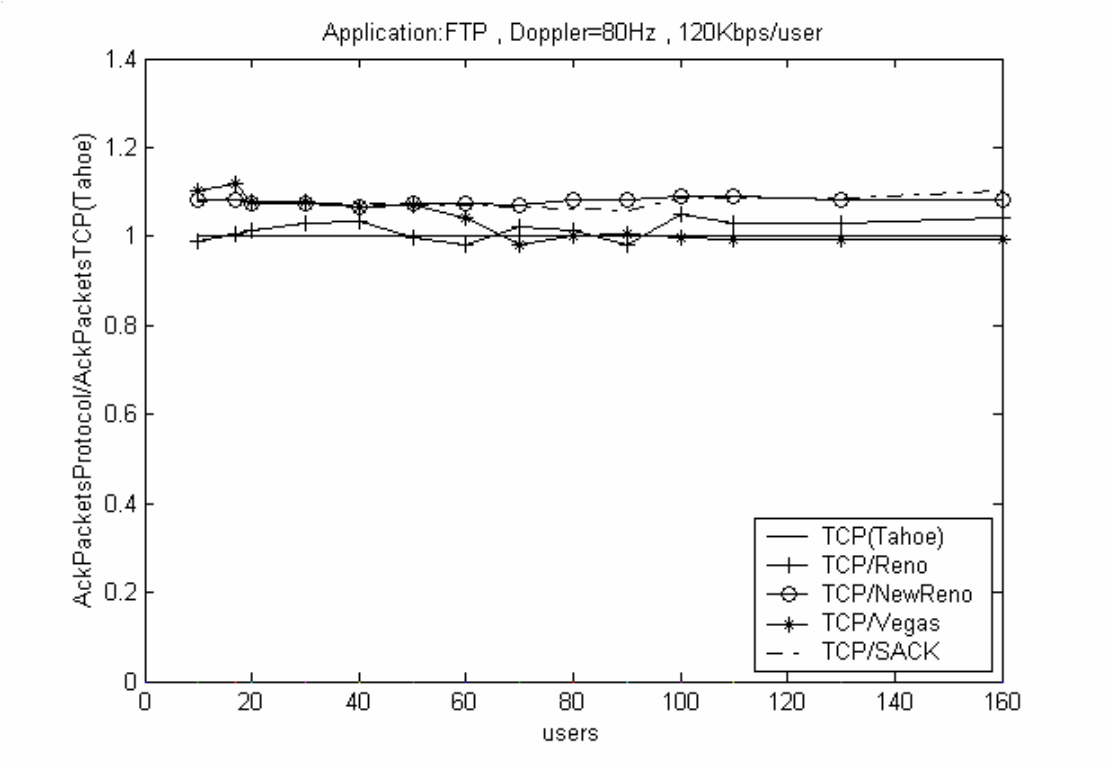
Graph 7. 2: Energy Consumption of FTP/6Hz/120Kbps



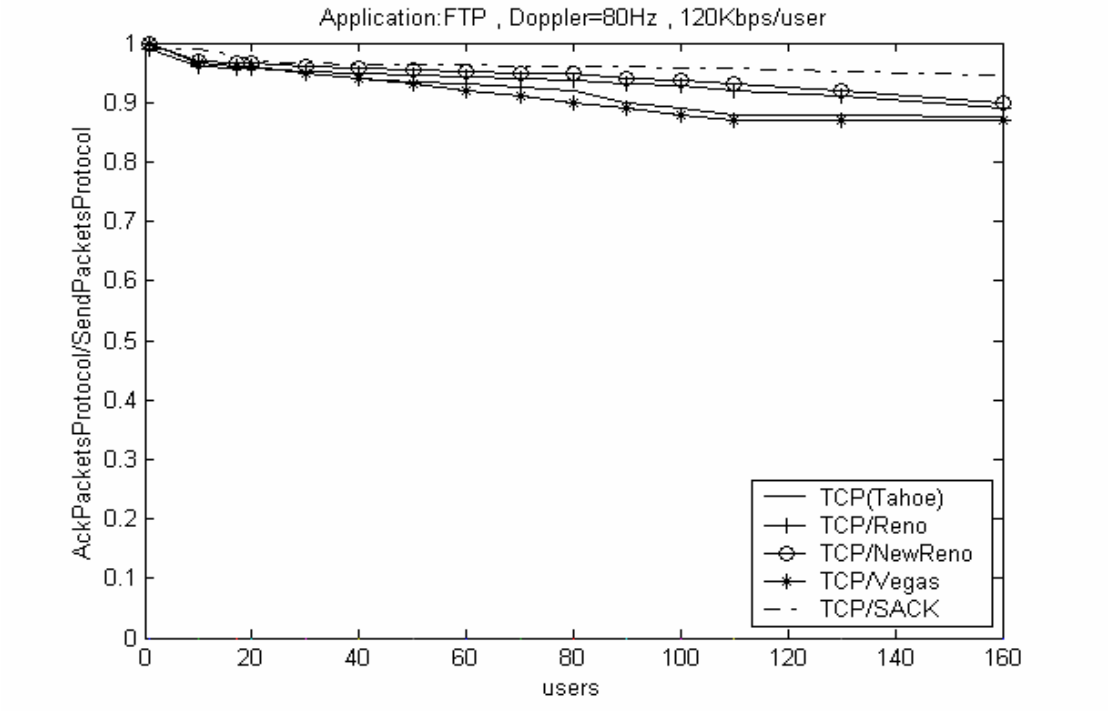
Graph 7. 3: Performance of FTP/20Hz/120Kbps



Graph 7. 4: Energy Consumption of FTP/20Hz/120Kbps



Graph 7. 5: Performance of FTP/80Hz/120Kbps



Graph 7. 6: Energy Consumption of FTP/80Hz/120Kbps

FTP application results show that TCP Vegas outperforms when users are less than 50 (in the case of 120Kbps/user) for f_D equal to 6 or 20 Hz and if users are less than 20 for f_D = 80 Hz. Beyond these numbers, the performance of the protocol deteriorates. This happens especially when the total numbers of users is more than 80 users. In this case TCP Vegas performance is the worst compared to the other types of TCP except TCP Tahoe. This is explained because TCP Vegas RTT estimation methodology can not be effective for heavy loads. On the other hand TCP New Reno and SACK are steadier for all Doppler frequencies. If f_D is equal to 6Hz, where a two-state Markov is used in order to model the errors, TCP Sack outperforms TCP New Reno slightly. When f_D is increased and is equal to 20Hz, a third state is added, a three-state Markov is used and as a result the frequency multi-packet loss in a single window increases. This is the reason why TCP outperforms even better than when f_D is equal to 6Hz [20]. If f_D is equal to 80Hz, the performance of TCP New Reno and TCP SACK is almost the same. Same results for the performance of TCP New Reno and TCP Tahoe over correlated high error rate environments agree with results already achieved in the literature [16] . Last but not least Reno outperforms Tahoe.

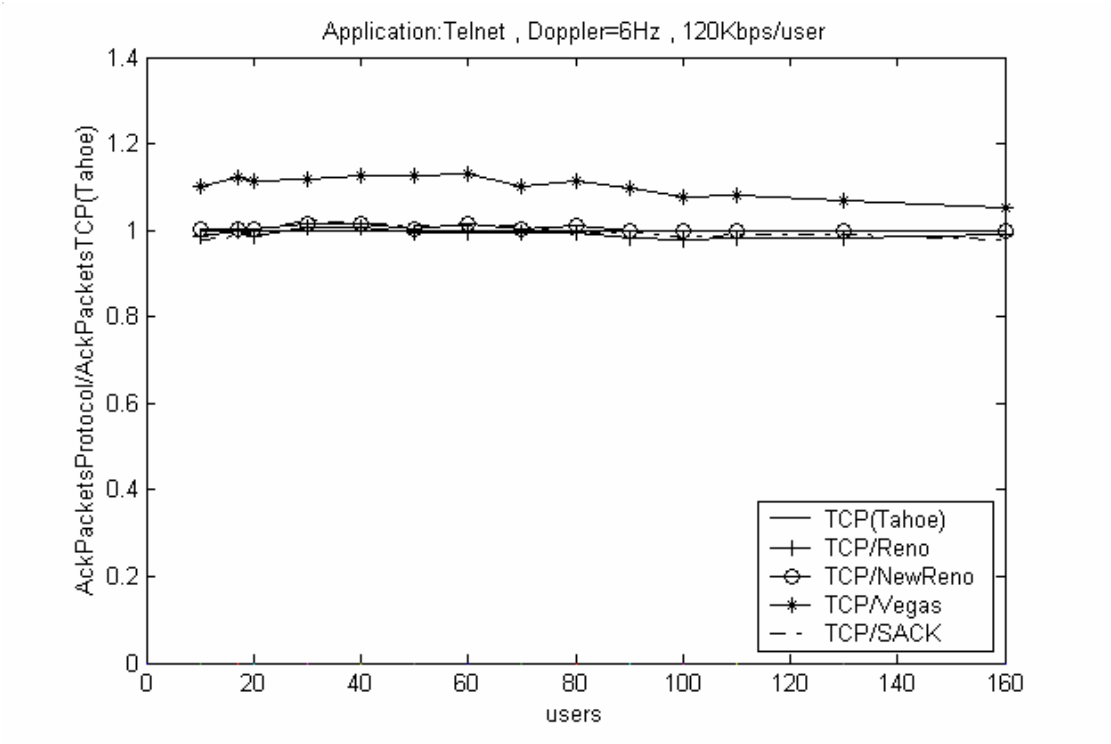
On the other hand TCP SACK consumes less energy than any other type of protocol because TCP SACK is adaptable [21]. This is illustrated in all cases. The reason why this happens is that TCP SACK has the ability to select which packets to retransmit. On the other hand TCP Vegas is the TCP version that consumes more than any other. TCP New Reno, TCP Reno and TCP Tahoe follow in that order.

Reduction of Energy Consumption is essential for UMTS because power control algorithms are utilized and because less battery power is then consumed. Moreover less CPU and memory capacity is used. Finally if the fraction of AckedPackets/SentPackets is utilized, then client's billing will be fairer especially in view of the fact that in UMTS the billing policy should be load oriented.

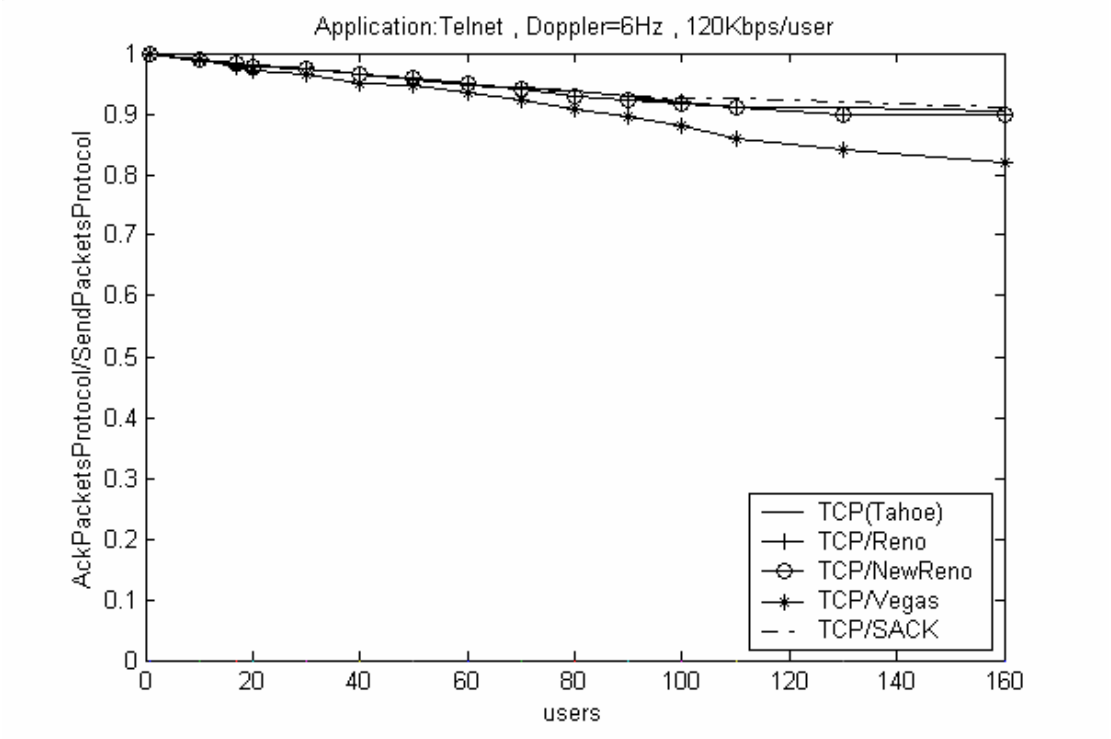
7.1.2 Telnet Application

As we will show in the following figures, the results for the Telnet application exhibit not many similarities with the corresponding results of the FTP application. TCP Vegas outperforms all others for f_D = 6 , 20 and 80Hz. . This happens because of the Telnet behavior. The Telnet client receives feedback as a result the estimation of RTT uses more samples and therefore is more accurate. The performance of all other TCP versions is similar

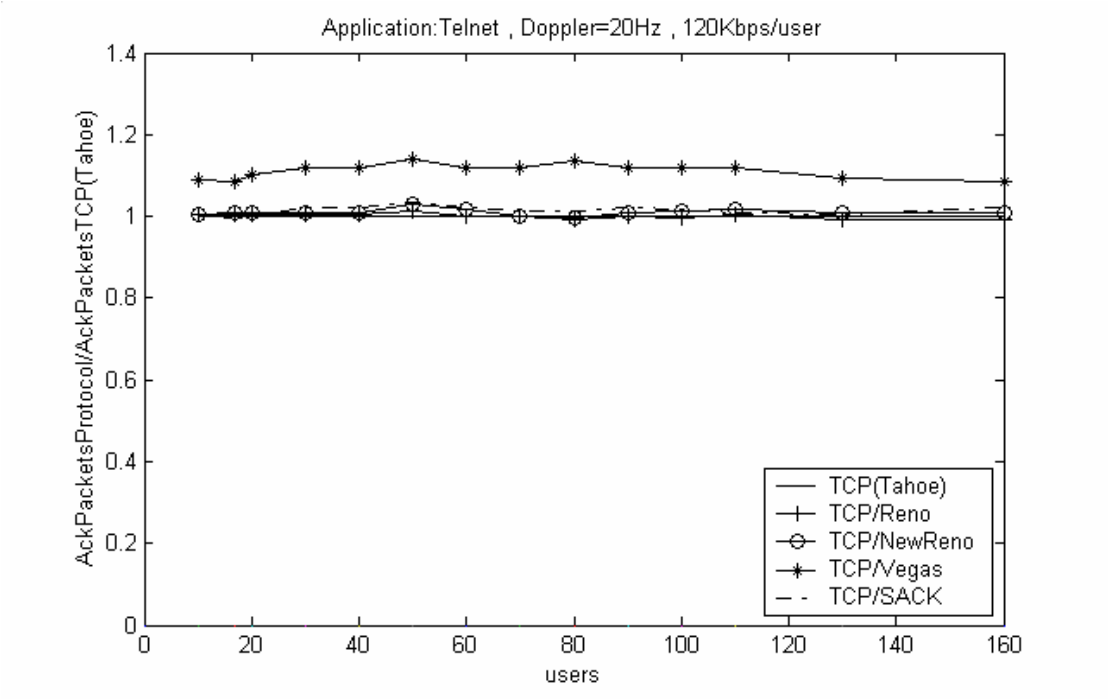
Concerning Energy Consumption, the behavior of the TCP protocol in the Telnet case follows that in the FTP behavior but now the deviation in larger even for small values of Doppler frequency. This means that energy consuming protocols, as was illustrates in the FTP application, consume more energy than used to in FTP, especially when f_D increases.



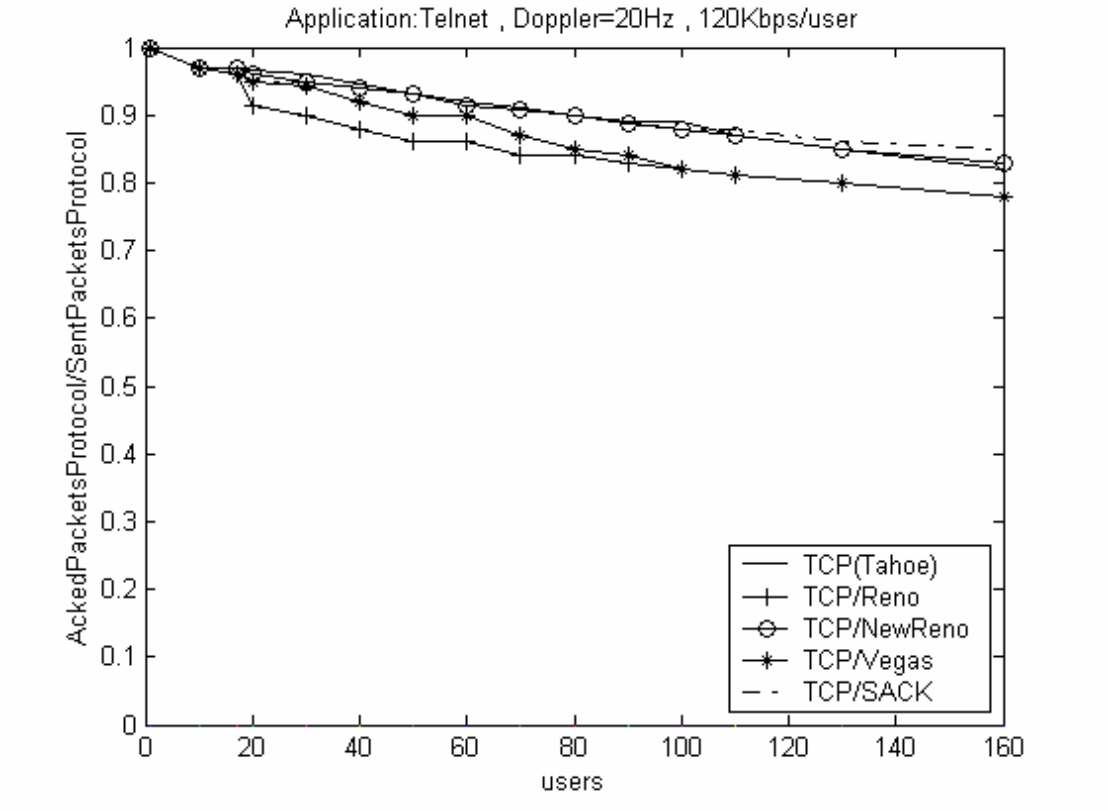
Graph 7. 7: Performance of Telnet/6Hz/120Kbps



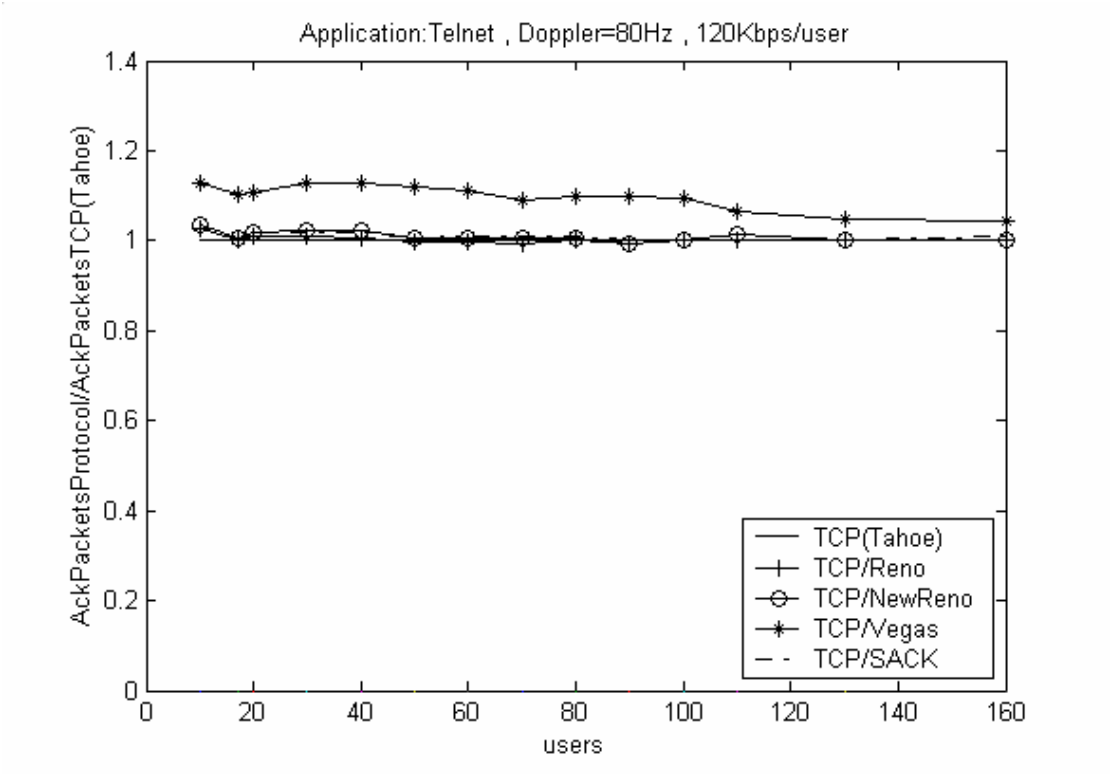
Graph 7. 8: Energy Consumption of Telnet/6Hz/120Kbps



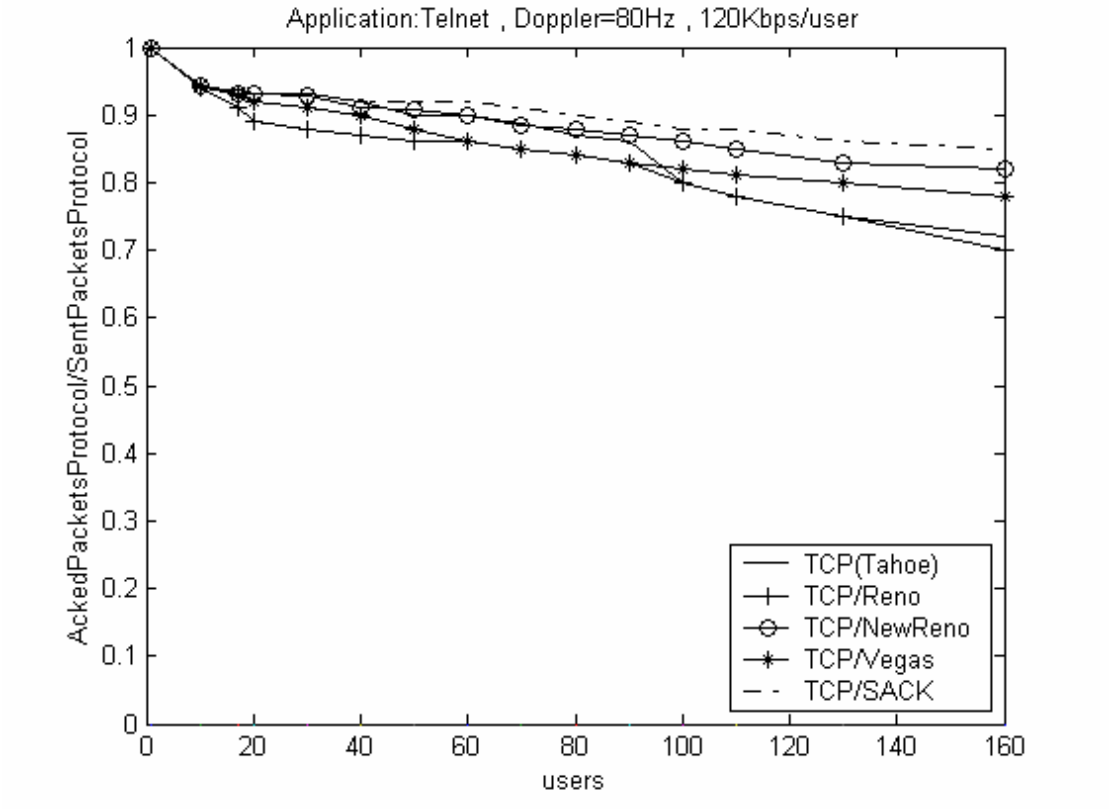
Graph 7. 9: Performance of Telnet/20Hz/120Kbps



Graph 7. 10: Energy consumption of Telenet/20Hz/120Kbps

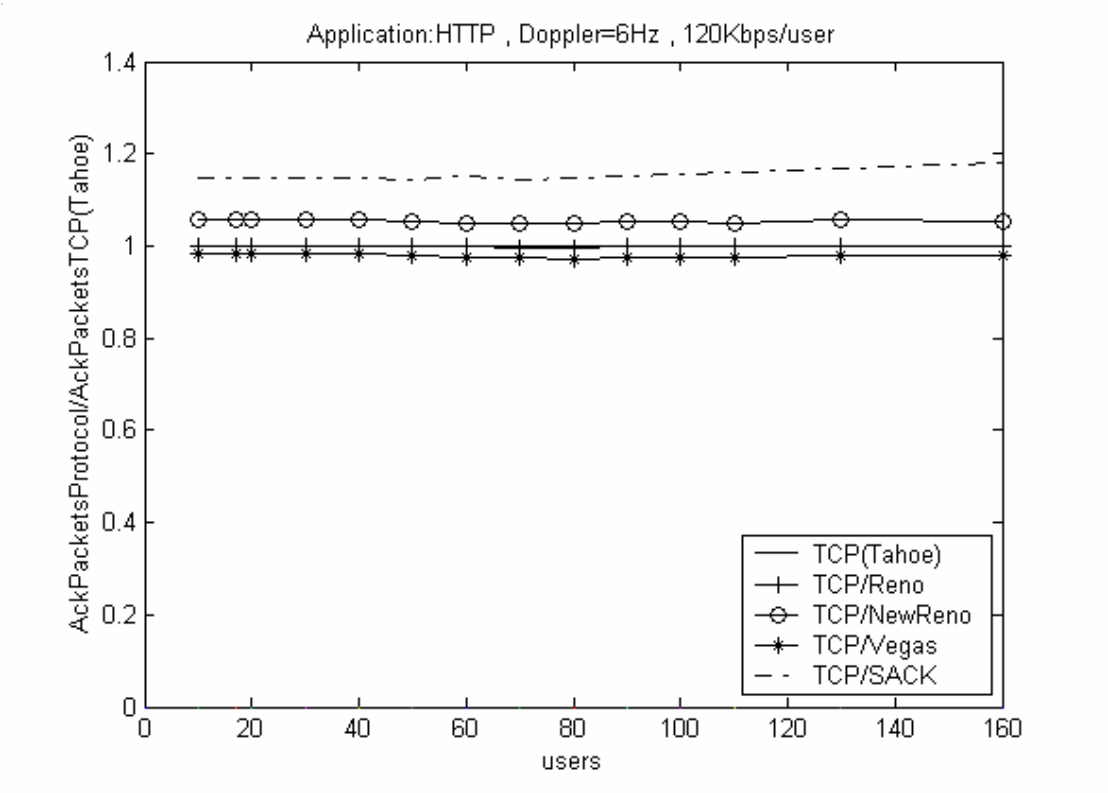


Graph 7. 11: Performance of Telnet/80Hz/120Kbps

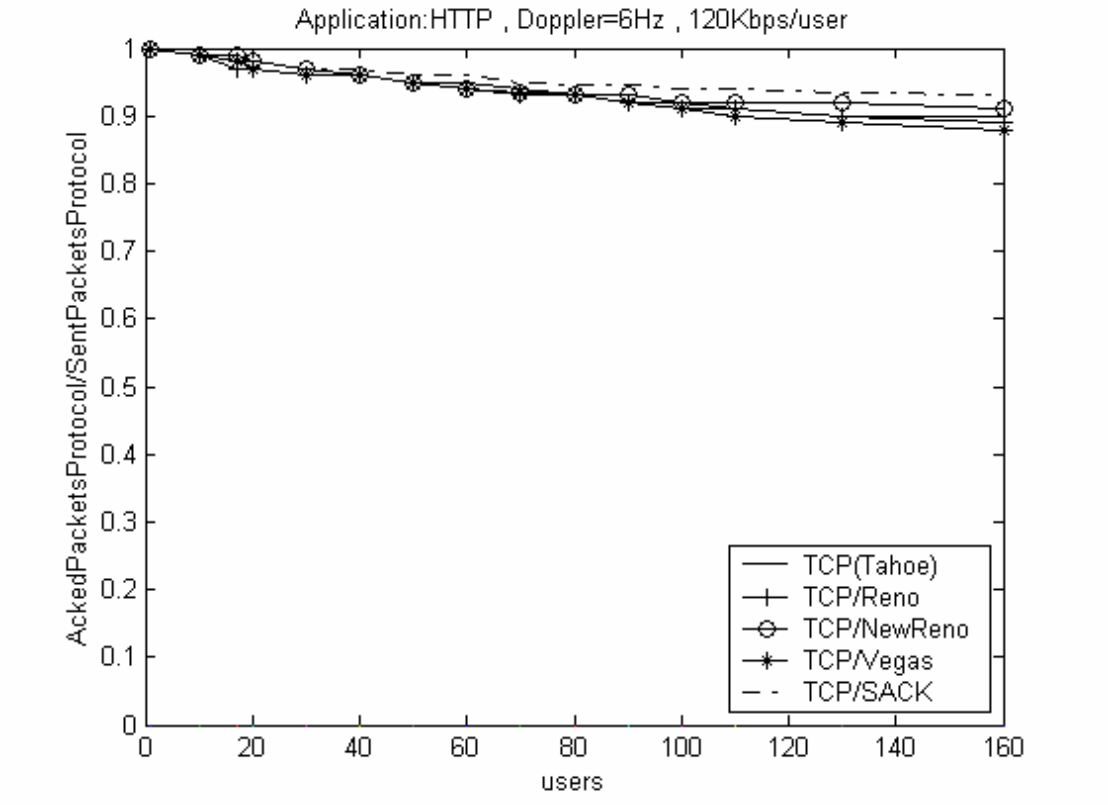


Graph 7. 12: Energy consumption of Telnet/80Hz/120Kbps

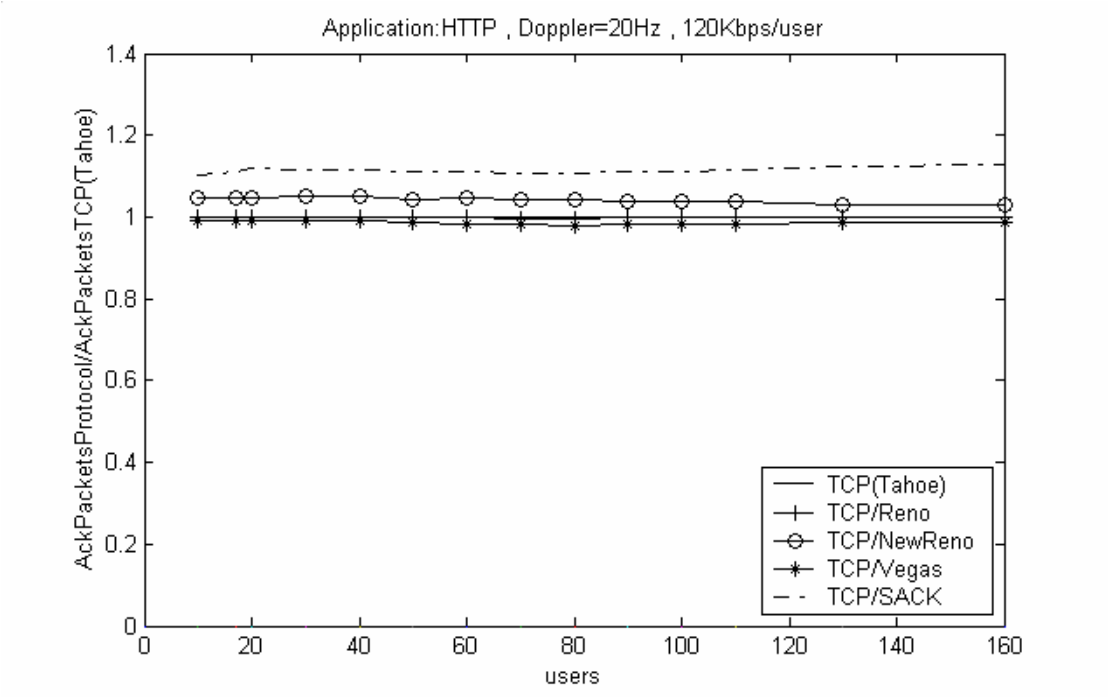
7.1.3 HTTP Application



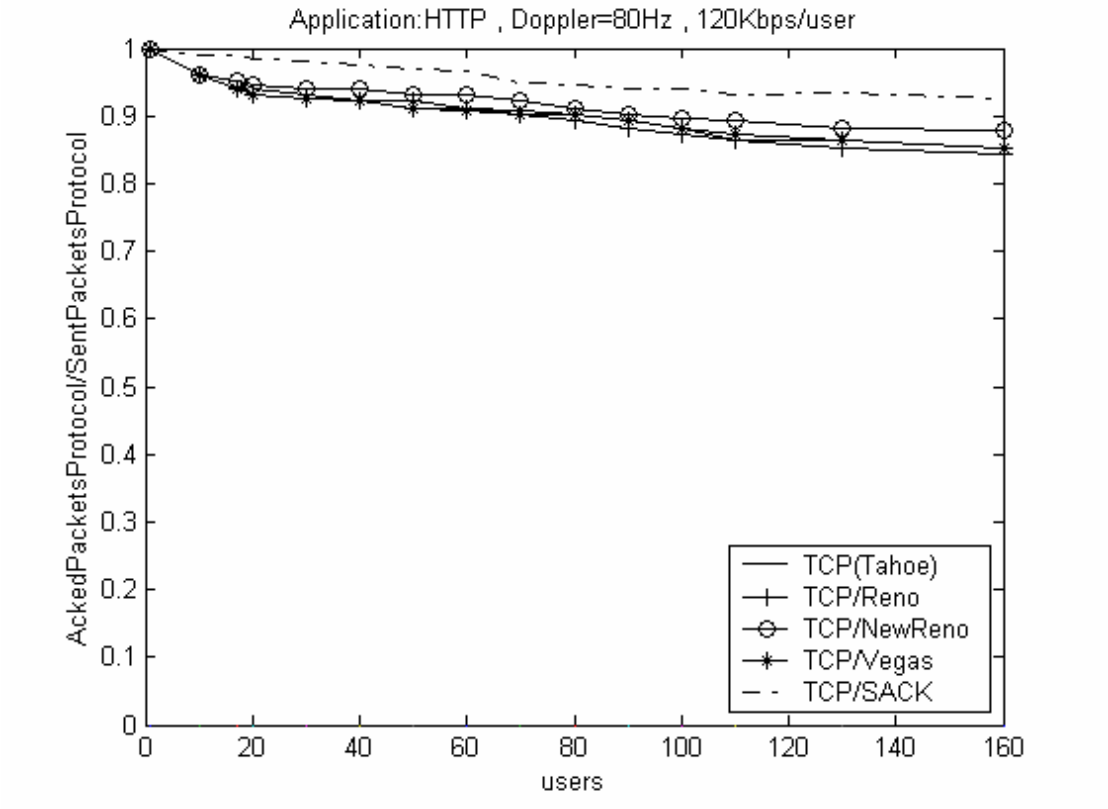
Graph 7. 13: Performance of HTTP/6Hz/120Kbps



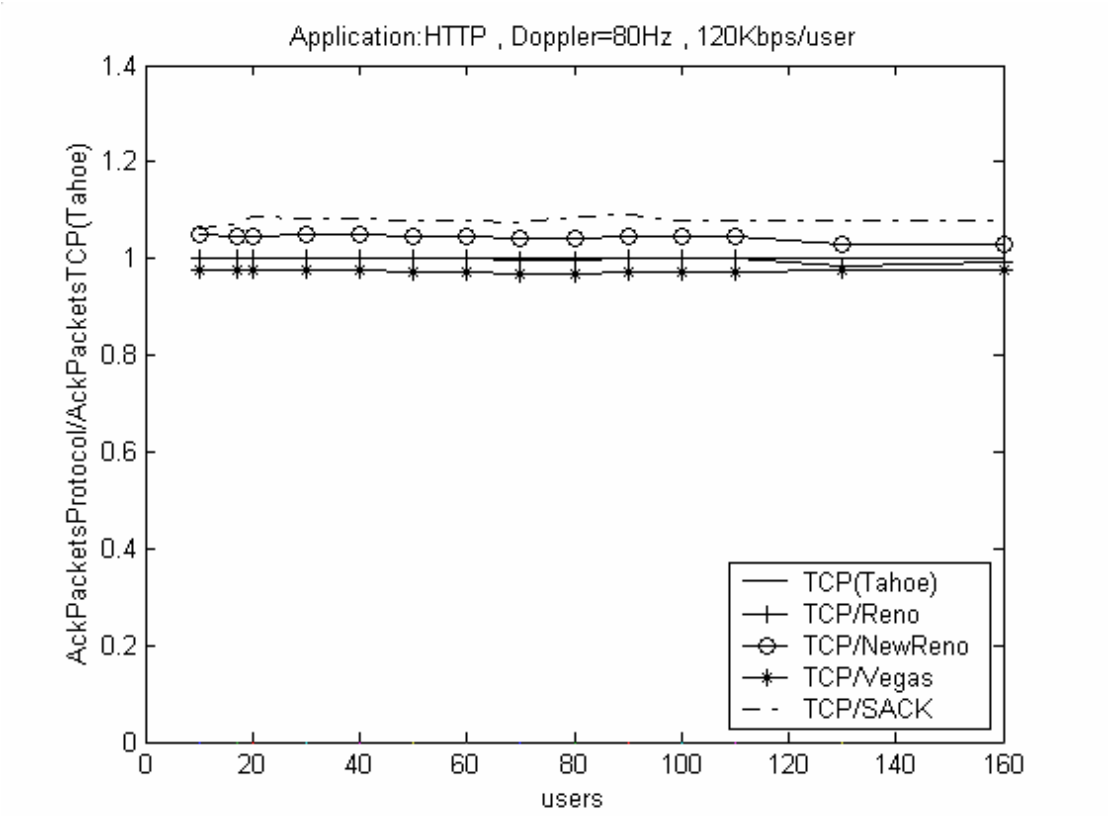
Graph 7. 14: Energy consumption of HTTP/6Hz/120Kbps



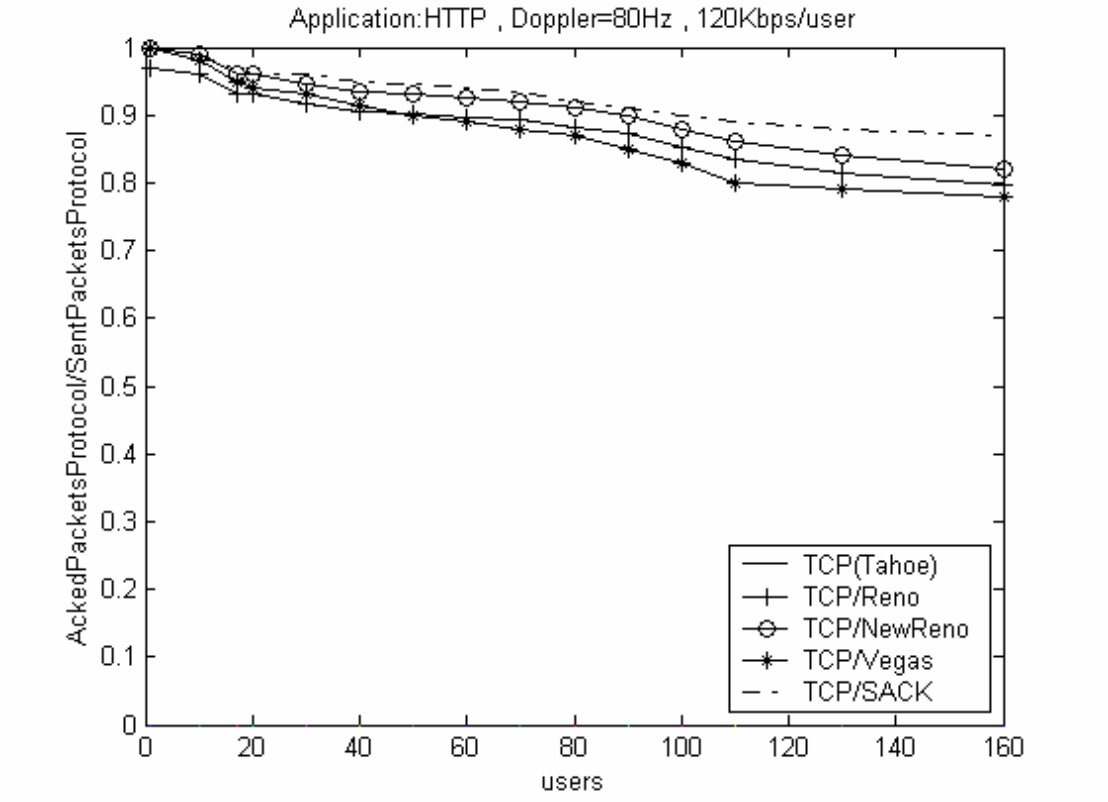
Graph 7. 15: Performance of HTTP/20Hz/120Kbps



Graph 7. 16: Energy Consumption of HTTP/20Hz/120Kbps



Graph 7. 17: Performance of HTTP/80Hz/120Kbps



Graph 7. 18: Energy Consumption of HTTP/80Hz/120Kbps

In the HTTP Application, TCP SACK outperforms in all cases and performs better as f_D increases. TCP New Reno performs better than TCP Reno which performs better than TCP Tahoe. TCP Vegas is the worst type because RTT estimation mechanism can not be adapted due to low load without feedback and mixed RTT behavior of HTTP (web or/and proxy).

Concerning Energy Consumption, TCP SACK performs even better and deviation between protocols is even larger.

7.1.4 Statement for users with 240Kbps bit rate

In the case of 240Kbps users we noticed the same behavior with 120Kbps link capacity in all applications and similar results about energy consumption have been observed. The only difference worth mentioning is that in the case of 240Kbps in HTTP application, TCP SACK still outperforms the others but to a smaller extend than before.

7.1.5 Proposed Approach

In future mobile devices for UMTS special operating systems will be installed. As has already be mentioned at the edges of UMTS (UE,BS,RNC) there will be a server/client agent. This agent will run over TCP. By selecting the best TCP version for each application without changing anything else (research results [26] have shown that this is the best solution) users will receive better TCP services. As explained above TCP Vegas outperforms in the cases of FTP with low load and Telnet. On the other hand TCP SACK outperforms in HTTP and FTP except the case of Doppler Frequency equal to 80Hz where TCP SACK has the same performance with TCP New Reno. Research results [15] indicate that over 75% of UMTS load will be HTTP related and less than 8% will be due to FTP and Telnet, as a result the selection of TCP SACK as a default TCP type is the right way. Furthermore, TCP SACK is the energy-efficient type of TCP. Moreover TCP New Reno would be the best alternative solution among classic TCP versions.

Most of current operating systems implement only one type of TCP. As this will change, then applications like FTP with low network load and Telnet can choose to run over TCP Vegas instead of running over TCP SACK.

We mention here that this proposal is based our simulation results for the UMTS Network standard of 3GPP. Possible further enhancements that will be proposed have not taken into account and may change the behavior of various versions of TCP [21].

7.2 Efficiency of TCP Applications

From the results presented in the figures it is clear that the performance of all types of TCP deteriorates over the lossy wireless UMTS which agrees with the literature [17],[18]. In order to provide a complete evaluation of TCP performance over UMTS, in the following section we will present efficiency of TCP results. We

have repeated our simulations without inserting any transmission errors. We define Efficiency of an Application as the fraction of:

TotalPacketsSentProtocl/TotalPacketsSentProtocol(no errors)

Our results agree with those reported in [5]. Reference [5] describes the efficiency of TCP New Reno over UMTS and claims that in a specific application, efficiency of each user depends on his average packet error rate and TCP is fair protocol, as already mentioned on chapter 5. In the next two tables we give the average Efficiency for FTP,Telnet and HTTP with Dopper frequency 6 , 20 and 80Hz for both 120 and 240 Kbps/user bit rate.

			Doppler (In Hz)		
TCP Type	BitRate/User (Kbps)	Application	4	20	80
TCP (Tahoe)	120	FTP	0.80	0.40	0.26
TCP Reno	120	FTP	0,85	0.41	0.26
TCP NewReno	120	FTP	0,87	0.44	0.27
TCP Vegas	120	FTP	0,83	0.42	0.23
TCP SACK	120	FTP	0,88	0.45	0.28
TCP (Tahoe)	120	Telnet	0,96	0.94	0.92
TCP Reno	120	Telnet	0,97	0.95	0.93
TCP NewReno	120	Telnet	0.97	0.95	0.93
TCP Vegas	120	Telnet	0.97	0.96	0.93
TCP SACK	120	Telnet	0.97	0.96	0.93
TCP (Tahoe)	120	HTTP	0.93	0.70	0.50
TCP Reno	120	HTTP	0.92	0.71	0.52
TCP NewReno	120	HTTP	0,93	0.72	0.53
TCP Vegas	120	HTTP	0.90	0.70	0.50
TCP SACK	120	HTTP	0.95	0.75	0.55
TCP (Tahoe)	240	FTP	0.76	0.37	0.24
TCP Reno	240	FTP	0,80	0.38	0.24
TCP NewReno	240	FTP	0,80	0.37	0.25
TCP Vegas	240	FTP	0,77	0.37	0.21
TCP SACK	240	FTP	0,82	0.40	0.28
TCP (Tahoe)	240	Telnet	0,95	0.94	0.92
TCP Reno	240	Telnet	0,97	0.95	0.92
TCP NewReno	240	Telnet	0.96	0.95	0.93
TCP Vegas	240	Telnet	0.97	0.95	0.93
TCP SACK	240	Telnet	0.97	0.96	0.92
TCP (Tahoe)	240	HTTP	0.88	0.69	0.48
TCP Reno	240	HTTP	0.90	0.71	0.50
TCP NewReno	240	HTTP	0,91	0.72	0.50
TCP Vegas	240	HTTP	0.85	0.68	0.44
TCP SACK	240	HTTP	0.92	0.72	0.52

Table 7. 1: Efficiency of TCP Applications

From these results is clear that loss, heavily effects all types of TCP. A packet loss of 1% results about 10% of loss of efficiency for FTP and about 5% loss of efficiency for HTTP for 120Kbps users and even worst for 240Kbps users. This happens due to TCP's reaction (TCP face any packet loss as indication of congestion), which is very defensive. Moreover as the frequency of packet losses increases the efficiency of all applications decreases rapidly. This is another reason why the best TCP version has to be chosen.

CHAPTER 8: Conclusions and Future Work

8.1 Conclusions

In this thesis we have studied the behavior of five TCP versions: TCP Tahoe, TCP Reno, TCP New Reno, TCP Vegas and TCP SACK over UMTS. We studied the behavior of these versions over different applications running over (FTP, Telnet and HTTP).

From the results of our simulations it is clear that the performance of all versions of TCP deteriorates. Moreover there is not an optimal TCP version for all applications and especially in different conditions: different Doppler frequency, however there is an optimal TCP version under specific conditions. TCP SACK outperforms in the case of the HTTP application under all conditions and in the case of FTP application for Doppler frequency equal to 6 and 20 Hz when the load is not very light. TCP Vegas outperforms in the Telnet application under all conditions and in the FTP application when there is very light load. In addition, although TCP New Reno does not outperform in any application, it is very steady.

On the other hand there is an optimal TCP version for Energy Consumption which is essential for the UMTS network. This version is TCP SACK (and TCP New Reno follows). Furthermore, TCP SACK outperforms in the HTTP application which will hold over 75% of UMTS load. This is why the UMTS network should use TCP SACK. Last but not least, TCP New Reno's can be an alternative solution.

One month ago, when our simulations have been completed, an IETF Draft ("TCP over 2.5/3G wireless networks" [23]) came out of the blue. One of main proposals was:

"TCP over 2.5G/3G SHOULD support SACK. In the absence of SACK feature, the TCP should use New Reno"

This proposal is in clear agreement with the results of our Thesis.

8.2 Future Work

Today, most of real time applications run over UDP and only application control runs over TCP. In the near future, new real time applications and protocols will be introduced with extensions adapted to error loss for special, use like streaming over UMTS. These protocols will basically run over TCP. An example is future RTSP. Future real time protocols will use hierarchical layers. Servers will send the basic layers of the information over TCP and only changes over UDP. If we model the traffic of this special TCP application then we can find the appropriate TCP version to use. Since we have already created a UMTS simulator, we can insert real-time

applications users with the above specifications and study the behavior of various TCP versions.

BIBLIOGRAPHY

[1] F. Muratore et al., "*UMTS : Mobile Communications for the Future*", John Wiley , 2001.

[2] B. Lieve and L. Suresh , "*Towards an All-IP-Based UMTS System Architecture*" , IEEE Network , January/February 2001.

[3] C. Anderson, "*GPRS and 3G Wireless Applications*", John Wiley, 2001.

[4] Bernhard H. Walke, "*Mobile Radio Networks – Networking and Protocols*" John Wiley, 1999.

[5] M. Zorzi,M. Rossi,G. Mazzini, "*Throughput and energy performance of TCP on a Wideband CDMA air Interface*" , WCDMA for UMTS : radio access for third generation mobile communications , John Wiley , 2000.

[6] M. Zorzi ,G. Mazzini, V. Tralli, A. Giovanardi, *"Some results on the error statistics over a WCDMA air interface"* , in Proc. MMT2000, Florida(USA), Dec. 2000.

[7] K.Heiska, "WCDMA Overview", Nokia Co.(Finland) Presentation, January 2002.

[8] J.Lahteenmaki, "*Radio Network Planning – Methods for Next Generation Systems*", Optimizing Next Generation Mobile Networks – ICM Conference, March 2000.

[9] P.M.Garrosa, "*Interactions between TCP and Channel Type*", Master of Science Thesis, Chalmers University of Technology, January 2002.

[10] V. Tsaousidis and I. Matta, *"Open Issues on TCP for Mobile Computing",* The Journal of Wireless Communications and Mobile Computing,Wiley Academic Publishers , Issue 2, Vol. 2, March 2002.

[11] H.Kaaranen, A.Ahtiainen, L.Laitenen, S.Naghian, V.Niemu, "*UMTS Networks – Architecture, Mobility and Services*", John Wiley, 2001.

[12] J. F. Kurose, K. W. Ross, *"Computer Networking: A Top-Down Approach Featuring the Internet"*, Addison Wesley, 2001.

[13] A. Lahanas and V. Tsaoussidis, *"Behavior of TCP-Probing with Hand-offs",* Internation Conference on Internet Computing, CSREA Press, Las Vegas, June 2001.

[14] S. Aust, N. Fikouras and C. Gorg, "Mobility Management in Integrated Network Platforms", ITG-Fachgruppe 5.2.4, June 2001.

[15] A. Klemm, C. Lindemann and M. Lohmann, "*Traffic Modelling and Characterization for UMTS Network*", GlobeCom2001, November 2001.

[16] F. Anjum and L. Tassiulas, "*An Analytical Model for the Various TCP Algorithms Operating Over a Wireless Channel*", IEEE WCNC'99, September 1998.

[17] T. Schwade and J. Schuler, "*Investigations on TCP Behavior during Handoff*", ITG Workshop Wurzburg, July 2001.

[18] J. Schuler and S. Gruhl, "Investigations on TCP traffic in mobile networks", ITG Workshop Bremen, January 2001.

[19] H. Balakrishnan, V. Padmanabhian, S. Sechan and R. Katz, "*A Comparison of Mechanisms for Improving TCP Performance over Wireless Links*", ACM SIGCOMM 1996, August 1996.

[20] K. Fall and S. Floyd, "*Simulation-based Comparisons of Tahoe, Reno and SACK TCP*", Computer Communication Review, July 1996.

[21] H. Singh and S. Singh, "*Energy Consumption of TCP Reno, Newreno, and SACK in Multi-Hop Networks*" , ACM SIGMETRICS 2002 , June 2002.
[22] D. Zhang, R. Zhang, Z. Kan, R. Cuny, J. Ruutu, J. Ma, "TCP over 2.5 and 3G wireless networks", Nokia Co. Literature review.
[23] H. Inamura, G. Montenegro, R. Ludwig, A. Gurtov, and F. Khafizov, "*TCP over second (2.5) and Third (3G) Generation wireless networks*", IETF draft, http://www.ietf.org/internet-drafts/draft-ietf-pilc-2.5g3g-10.txt
[24] Network Simulator Manual, http://www.isi.edu/nsnam/ns/doc/ns_doc.ps.gz
[25] M. Milani, G. Rosso, "*Internet Applications and protocols performance in UMTS mobile Networks*", XIV Master in Information Technology, July 2002.