Link-layer error recovery techniques to improve TCP performance over wireless links

Claudia Rinaldi

February 25, 2005

Contents

Abstract

Recent technology has involved TCP in wireless applications even if it was originally designed to work over wired links. It is well known that wireless links are usually characterized by phenomena like shadowing, multipath propagation and path loss that cause an increase in the bit error rate. Moreover, TCP protocol was supposed to interpret all the losses as due to congestion because of it was designed to work over wired networks. All these problems caused by the wireless application of TCP usually degrade TCP performances. In order to improve TCP behavior over wireless links several solutions have been proposed in literature as: the split connection approach, the end-to-end approach and the link-layer approach. Starting from the link-layer solution that have been studied by Barman and Matta, the main contribution of this thesis is the study of the effects of different modulation formats on the maximum achievable value of an *objective function*, defined as the ratio between the TCP throughput and a cost function. Appropriate power management and error correction techniques are assumed to improve the link reliability observed by TCP and increase the objective function performance accordingly.

Chapter 1

Introduction

TCP provides a reliable transport service to many of today's Internet applications and it has been thought out, designed and optimized for wired networks. That is why it has severe performance problems when wireless links are involved in the end-to-end connection as a result of the introduction of mobile radio systems. In fact TCP reacts to packet losses by initiating a congestion control or avoidance mechanism and by backing off its retransmission timer (Karns algorithm).
These measures result in a reduction in the load on the intermediate links, thereby controlling congestion in the network. Unfortunately, when packets are lost in networks for reasons other than congestion (e.g. multipath fading, atmospheric conditions, user mobility...) these measures result in an unnecessary reduction in end-to-end throughput and hence in suboptimal performance.
A lot of solutions have been proposed in the last years in order to solve these critical aspects.

The aim of this thesis is to focus on the link layer approach, in particular using both FEC and ARQ to solve the problem that TCP interprets all the losses as due to congestion. The effect of the amount of the bandwidth consumed by both FEC redundancy and different modulations used is also considered in order to get the maximum gain in TCP performance.

The outline of the thesis is as follows: in chapter 2 an introduction to the general environment of the simulations is presented, chapter 3 is related to the problems caused by the application of TCP over wireless links while the previously proposed solutions to solve these problems are summarized in

chapter 4, after the definition of the problem, the model used in the simulations is described in chapter 5, the definition of the function to be optimized is given in chapter 6 and in chapter 7 the simulations and results obtained are presented, finally some ideas for future works and the conclusions are given respectively in chapters 8 and 9, moreover two appendix follow the conclusions in order to clarify some technical aspects.

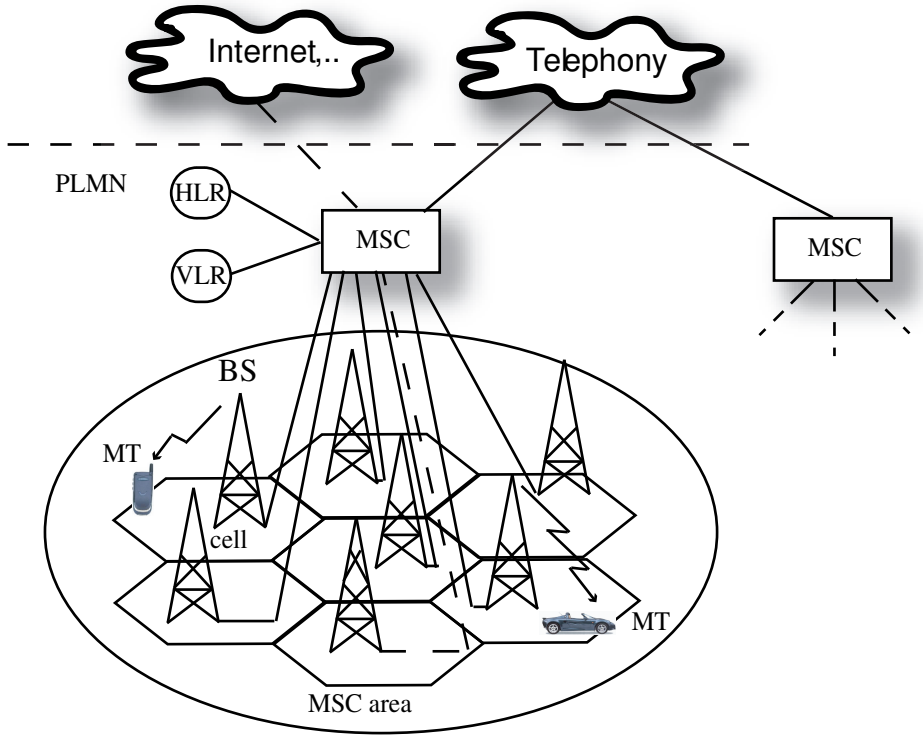
Chapter 2

Background

2.1 Architecture of a general cellular network

The general architecture of a wireless/wired system is shown in figure 2.1 where the main interest has to be given to the mobile terminal (MT) and the base station (BS) that communicate through the use of a wireless link. This is an example of a radio mobile transmission because it is usually defined as a link between two terminals, where at least one of them is moving. As previously said, the main interest of this thesis is on the application, in a wireless environment, of a protocol (TCP) that was designed to work in wired links, this is for instance the case of the mobile terminal that wants to connect to the internet. In order to understand the background that as been analyzed a short description of problems related to the use of a mobile radio system is following.

A mobile radio system is a network of linked base stations with the aim of furnishing the radio coverage of a certain service area. Land-mobile communication is burdened with particular propagation complications compared to the channel characteristics in radio systems with fixed and carefully positioned antennas. The antenna height at a mobile terminal is usually very small, typically less than a few meters, so obstacles and reflecting surfaces in the vicinity of the antenna have a substantial influence on the characteristics of the propagation path. Moreover, the propagation characteristics change from place to place and, if the mobile unit moves, from time to time. In generic system studies, the mobile radio channel is usually evaluated from *statistical* propagation models: no specific terrain data is considered, and



MT: mobile terminal
MSC: mobile switching center
VLR: visitor location register
HLR: home location register
PLMN: public land mobile network
PSTN: public switched telephone network

Figure 2.1: Cellular Transmission Network.

channel parameters are modelled as stochastic variables.
Three mutually independent, multiplicative propagation phenomena can usually be distinguished [8]: multipath propagation, shadowing and large-scale path loss.

- *Multipath Propagation*: it is caused by a multiple reception of the same signal due to different obstructions that this signal meets in the path from the sender to the receiver. This leads to rapid fluctuations of the phase and amplitude of the received signal if the vehicle moves over a distance in the order of a wave length or more (a few meters). Multipath fading thus has a *small-scale* effect.

- *Shadowing*: it is a *medium-scale* effect. Field strength variations occur if the antenna is displaced over distances larger than a few tens or hundreds of meters.

- *Large Scale Path Loss*: these cause the received power to vary gradually due to signal attenuation determined by the geometry of the path profile in its entirety. This is in contrast to the local propagation mechanisms, which are determined by building and terrain features in the immediate vicinity of the antennas.

2.1.1 Multipath Propagation

In typical link between a BS and a MT there are a lot of scatterers, that is why the signal offered to the receiver contains not only a direct *line of sight* radio wave but also a large number of reflected radio waves. These reflected waves interfere with the direct wave, which causes significant degradation of the performance of the network. A wireless network has to be designed in such way that the adverse effect of these reflections is minimized. Although channel fading is experienced as an unpredictable, stochastic phenomenon, powerful models have been developed that can accurately predict system performance: narrow band Rayleigh, or Rician models mostly address the channel behavior at one frequency only; dispersion is modelled by the delay spread. The effects of multipath reception are summarized in table 2.1.

Fast moving user	Rapid fluctuations of the signal amplitude
Wideband digital signal	Dispersion and Intersymbol interference
Analog television signal	ghost images (shifted slightly to the right)
Stationary user of a narrowband system	Good reception at some locations and frequencies; poor reception at other locations and frequencies
Satellite positioning system	Strong delayed reflections may cause a severe miscalculation of the distance between user and satellite

Table 2.1: Effects of multipath reception

2.1.2 Shadowing

Experiments reported by Egli in 1957 showed that fluctuations in the received power caused by the variation of environmental conditions can be represented by a log-normal distribution. This log-normal model furnishes a description to the first order of the shadowing phenomena. *Log Normal* means that the local power expressed in logarithmic values has a Gaussian distribution, thus, the probability density function of the local mean power is in the form:

$$f_{p_{log}} = \frac{1}{\sqrt{2\pi\sigma_s^2}} \exp\left[-\frac{1}{2\sigma_s^2} p_{log}^2\right] \tag{2.1}$$

where σ_s is the logarithmic standard deviation of the shadowing, expressed in natural units.

2.1.3 Large Scale Effects: Path Loss

There are a lot of empirical model to predict the mean power level of the received signal as the function of some system parameters and for different environments. The most frequently used model is the one of Okumura-Hata that furnishes an estimation of the mean power attenuation as a function of the distance from the base station, the frequency of the radio link, the height of both antennas (the one of the base station and the other one of the mobile user). One example is shown in figure 2.2, where [9]:

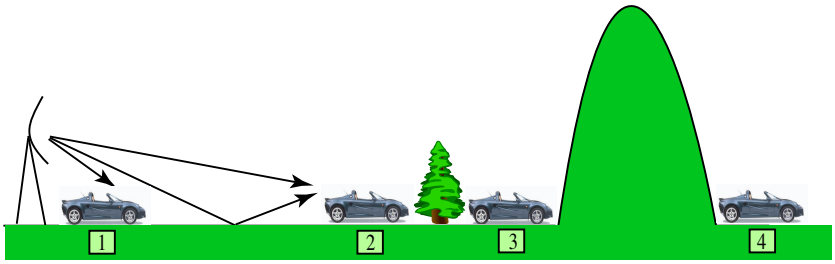


Figure 2.2: Components influencing the attenuation

- Mobile 1 is subjected to free space loss.
- A strong line-of-sight is present for mobile 2, but ground reflections can significantly influence path loss. The plane earth loss model appears appropriate.
- For mobile 3 plane earth loss needs to be corrected for significant diffraction losses, caused by trees cutting into the direct line of sight.
- For mobile 4 loss prediction is fairly difficult and unreliable since multiple diffraction is involved.

2.2 Introduction to TCP/IP

The presented work is related to the Transport Layer of the TCP/IP protocol stack. All modern networks are designed using a layered approach, where each layer presents a predefined interface to the layer above it. The ISO/OSI protocol with seven layers is the usual reference model. Since TCP/IP was designed before the OSI reference model it has only four layers [11], [12]. The aim of this section is to explain the general organization and implementation of modern computer networks, showing, in this environment, position and functionalities of TCP. A TCP/IP network is generally a heterogeneous network, meaning there are many different types of network computing devices attached. The suite of protocols that encompass TCP/IP were originally designed to allow different types of computer systems to communicate as if they were the same system. It was developed by a project underwritten by an agency of the Department of Defense known as the Advanced Research Projects Agency (DARPA). TCP/IP is a family of protocols. As all other

Internet Protocol Suite	
Application	**HTTP, SMTP, FTP, SSH, IRC, SNMP...**
Transport	**TCP, UDP, SCTP, RTP, DCCP...**
Network	**IPv4, IPv6, ARC, ICMP...**
Data Link	**Ethernet, 802.11 WiFi, Token Ring, FDDI...**

Figure 2.3: Internet Protocol Suite

communication protocols TCP/IP is composed by layers as shown in figure 2.3. The best way to introduce TCP/IP is by looking at it through the ISO OSI model and studying its protocols and functions basing on their placement in the OSI model as it has been done in figure 2.4.

The heart of the TCP/IP network protocol is at layers 3 and 4. The applications for this protocol (file transfer, mail, and terminal emulation) run at the session through the application layer. It is important to notice that this protocol runs independently of the data-link and physical layer. At these layers, the TCP/IP protocol can run on Ethernet, Token Ring, FDDI, serial lines, X.25, and so forth. It has been adapted to run over any LAN or WAN protocol.
Basing on the figure 2.4, a description of the different layers is following, giving much more attention to the network and the transport layers.

2.2.1 Link layer

The lowest layer of the OSI Reference Model is the physical layer, which is responsible for transmitting information from one place to another on a network. The layer just above the physical layer is the data link layer, called the network interface layer or just the link layer in the TCP/IP architectural model. Its primary job is to allow packets transmission between devices on the same physical link and to interface between the hardware-oriented physical layer, and the more abstract, software-oriented functions of the network

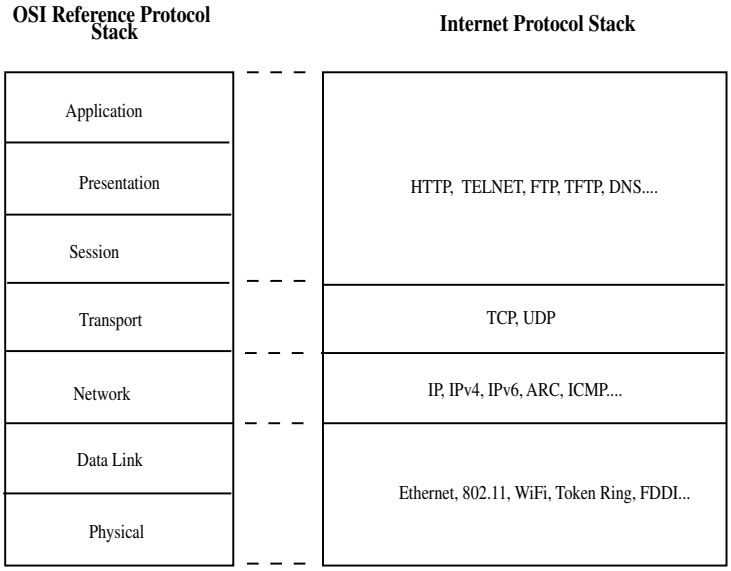


Figure 2.4: Comparison between OSI and TCP/IP protocol stacks

layer and those above it. The TCP/IP protocol suite is structured around the Internet Protocol (IP). IP operates at layer three of the OSI Reference Model, and assumes that it will be layered on top of an existing layer two technology. However, certain types of connections exist that do not include a layer two protocol over which IP can run. To enable TCP/IP to operate on these kinds of links, two special TCP/IP data link layer protocols have been created: the Serial Line Internet Protocol (SLIP) provides a layer two framing service for IP datagrams and the Point-to-Point Protocol (PPP) which defines a complete method for robust data link connectivity between units using serial lines or other physical layers. It includes numerous capabilities and features, including error detection, compression, authentication, encryption and much more.
Anyway, basing on the OSI reference model a general layer 2 has to assure the functionalities below:

- *Physical addressing* (as opposed to network addressing) defines how devices are addressed at the data link layer.
- *Network topology* consists of the data link layer specifications that often define how devices are to be physically connected, such as in a bus or a ring topology.

- *Error notification* alerts upper-layer protocols that a transmission error has occurred, and the sequencing of data frames reorders frames that are transmitted out of sequence.

- *Flow control* moderates the transmission of data so that the receiving device is not overwhelmed with more traffic than it can handle at one time.

2.2.2 Network layer

The network layer offers the same services of the corresponding third level of the ISO/OSI reference model. Moreover it defines a specific protocol for data transmission through non-homogeneous networks called Internet Protocol (IP) and other protocols for network control.

IP protocol

The IP is designed to interconnect packet switched communication networks to form an internet. It transmits blocks of data called *datagrams* received from the IP's upper-layer software to and from source and destination hosts and, to achieve this, it implements two functions: *addressing* and *fragmentation*. In order to allow for multiple IP networks to interoperate, there must be a mechanism to provide flow between the differently addressed systems. The device that routes data between different IP addressed networks is called a *router* and the protocols that distribute the IP address information to each router are called *routing protocols*.

IP offers a connectionless delivery service for the logical communication between two hosts. It means that this protocol does not set up a session (virtual link) between the transmitting and the receiving stations prior submitting the data to the receiving station. Moreover, it does not guarantee to successfully send all the datagrams between the two hosts, permitting lost of datagrams, errors in datagrams and reception of datagrams out of order. Whenever a loss occurs IP does not inform anyone and it is up to the upper layer protocols (TCP, or even the application itself) to perform error recovery.

2.2.3 Transport layer

The transport layer of the TCP/IP model is concerned with the same functionalities of the corresponding layer in the ISO/OSI reference model. In this overview only two protocols called Transport Control Protocol (TCP) and User Datagram Protocol (UDP) are described.

Transmission Control Protocol

TCP is a connection oriented transport service, the main services guaranteed by TCP are:

- Basic data transfer.
- Reliability.
- Flow control.
- Multiplexing.
- Connections.
- Congestion control.

TCP enables two hosts to establish a connection and exchange streams of data which are treated in bytes. The delivery of data in the proper order is guaranteed by the use of sequence numbers [1]. Moreover, through the use of both acknowledgements and sequence number, TCP can detect errors or lost data and can trigger retransmission until the data is received, complete and without errors.
Everything that TCP sends is called a *segment*. This informational unit can be control data or user data. A TCP segment will contain the TCP header, figure 2.5, and its data. The data handed to TCP for transmission is known as a stream; more specifically, an unstructured stream. When TCP receives a datastream from the application, it will divide the data into segments for transmission to the remote network station. A segment can have control or data information, it is simply an unstructured stream of data bytes sent to a destination.

[1] It has to be noted that basing on the particular way of assignment of sequence numbers TCP is said to be *byte oriented* because a sequence number is assigned to every byte in each packet.

When an application wants to communicate with another application via TCP it sends a communication request. This request must be sent to an exact address. After an *handshake* between the two applications, TCP will setup a *full duplex* communication between the two applications. TCP is responsible for breaking segments into IP packets before they are sent and for reassembling the packets when they arrive, furthermore it has to verify the correct delivery of data.

One function is required to TCP in order to manage a connection: *flow control*. Flow control is used in order to avoid the sender to fill the buffer of the receiver, in practise it has been designed to control the speed of the sender basing on the speed used by the receiver to read the incoming data. Flow control is achieved by the using of a sender's window, called *receive window*, that is periodically set to the value of the receivers buffer.

Another feature of TCP is the so called *congestion control* algorithm, which uses another window to the senders TCP called *congestion window*. This is not negotiated in the establishment of the connection neither advertised in the TCP header; it is assumed. Congestion controlled is composed by three main phases:

- Additive increase, multiplicative decrease: when a loss notified by a triple-duplicate ACK occurs, the value of the congestion window is halved[2] while, when no losses are notified, the growth of the congestion window is linear and this second phase is also known as *congestion avoidance*.

- Slow start phase: it is used at the establishing of a connection. During this phase the congestion window is initially set to 1 segment and, if loss events are not notified, the congestion window is exponentially increased.

- Reaction to time outs events: in this case a loss is notified by the expiration of a time out and the congestion window is reduced to 1 segment while entering the phase of slow start.

[2]This a particular characteristic of the more recent version of TCP called Reno, while in the previous version called Tahoe, the reaction to losses due to both triple-duplicate ACK and time out events was the reduction of the congestion window to 1.

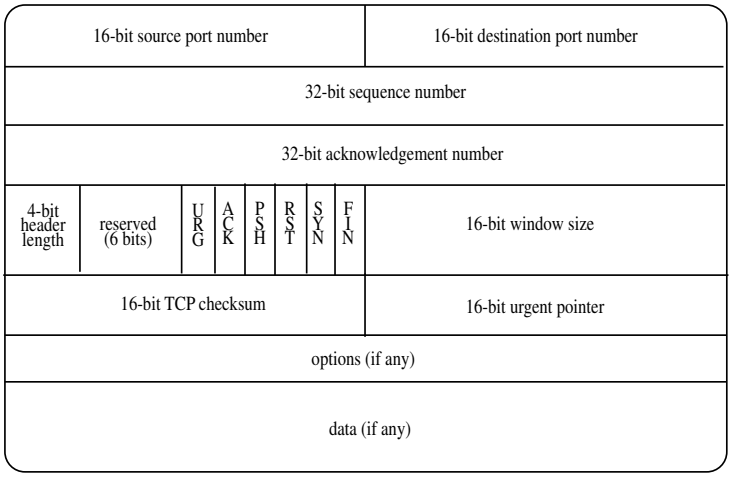


Figure 2.5: TCP Header

All these dynamics are controlled by TCP through the use of a variable called *slowstart threshold*. This variable establishes the value of the congestion window after which the phase of slow start has to finish and the phase of congestion avoidance has to begin.

User Datagram protocol

It is a connectionless, unreliable transport service. It does not issue an acknowledgment to the sender upon the receipt of data. It does not provide order to the incoming packets, and may lose packets or duplicate them without issuing an error message to the sender. This should sound like the IP protocol. The only offering that UDP has is the assignment and management of port numbers to uniquely identify the individual applications that run on a network station and a checksum for error detection. UDP tends to run faster than TCP, for it has low overhead (8 bytes in its header compared to TCP's typical 40 bytes). It is used for applications that need a real-time transport. Any application program that incorporates the use of UDP as its transport-level service must provide an acknowledgment and sequence system to ensure that packets arrive, and that they arrive in the same order as they were sent.

2.2.4 Application layer

The suite of applications specifically developed for TCP/IP protocol is the following, [18], [26], [27]:

- HTTP is the de facto standard for transferring World Wide Web documents, although it is designed to be extensible to almost any document format.
- TELNET runs on the top of the TCP protocol and allows a network workstation to appear as a local device to a remote device.
- File Transfer Protocol (FTP) allows data files to be reliably transferred on the Internet.
- Trivial File Transfer Protocol (TFTP) is based on the unreliable transport layer, UDP, and it is used for boot loading of configuration files across an Internet.
- Domain Name Service (DNS) allows users to establish a connection to network stations using human names instead of cryptic network addresses.
- Simple Mail Transfer Program (SMTP) is an electronic mail system that is robust enough to run in the entire internet system.
- Boot Protocol (BOOTP) and Dynamic Host Configuration Protocol (DHCP) allow for management of IP parameters on a network.

There are many other applications that run on a network using the TCP/IP protocol suite that are not shown here. Included in this listing are the applications that are defined in the RFCs and are usually included in every TCP/IP protocol suite that is offered. However, newer applications or protocols for TCP/IP are sometimes not included.

2.2.5 Differences between OSI protocol stack and TCP/IP protocol stack

TCP/IP and OSI are both based on the concept of independent protocol stacks but they are characterized by opposite approaches to this problem. The OSI model was in fact conceptualized first and then implemented, while

TCP/IP Reference Model was actually drawn up to describe the already existing network stack. The first evident difference between them is the number of layers, TCP/IP has in fact only four layers because it has not the session layer and the application and presentation layers are not separated as shown in figure 2.4. Moreover the OSI model makes clearly the differences between three main concepts: *services*, *interfaces* and *protocols* while this do not happen for the TCP/IP model. The last important difference comes from the mode of connection. Indeed, connection-oriented modes and connectionless modes are available in both models, but not at the same layer: in the OSI model, these modes are only available at the network layer (at the transport layer, only the connection-oriented mode is available), though they are available at the transport layer in the TCP/IP model (the internet layer only offers the connectionless mode). Therefore, the TCP/IP has an advantage, compared to the OSI model: applications (that directly use the transport layer) have the choice between both modes.
To conclude, the OSI Reference Model has proven to be invaluable in the realm of education because it was conceived on an abstract level. It offers a very clean model with clear distinctions between the roles of each layer. Unfortunately in practice things are often not so clear cut. For example, error correction may need to be done at each layer to satisfy its correctness, but in practice a higher layer to correct everything below it for efficiency's sake would be required. Another problem with OSI is that some layers are overkill. For instance, the session and presentation layers both do very little, and what they do is likely to vary from application to application. Thinking about them conceptually is useful, but implementing usually is not. The TCP/IP protocol stack, on the other hand, is not even a proper model. For one thing, only the middle two layers, network and transport are well-defined. The bottom layer mostly defines an interface to some heterogenous hardware below it. This is done so that TCP/IP can be implemented on top of any simple network, but it renders the model incomplete because it tells nothing about how data transmission is accomplished at the frame or physical levels.

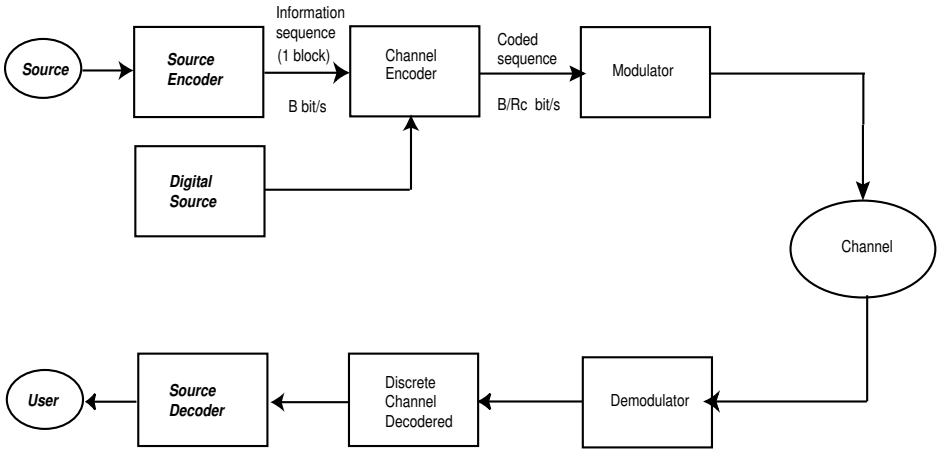


Figure 2.6: Transmission Scheme

2.3 Introduction to a general communication system

Since this thesis is related to many components and mechanism characterizing a general communication system, a short introduction of these topics is following.

The basic elements of a digital communication system are shown in figure, 2.6, where both the wired and the wireless links are supposed to be affected only by *white Gaussian Noise*, [4]. The message that enters the encoder is analog and it is produced by a so called *Analog Source*. Until the 1960s analog signals were directly transmitted over the channel through an *Analog Communication System*.

In a digital communication system, instead, analog signals are usually converted in a sequence of binary digits by a block called *encoder* or *source encoding*. Then, these digits has to be transmitted through a channel, but usual channels cannot be used for transmitting binary digits from the source because they are *waveform channel* (such as pair of wires, coaxial cables, optical fibers...). Therefore another block is added in the transmission scheme with the aim of converting binary digits into waveforms compatible with the channel characteristics. This block is called *Digital Modulator* or, simply, *modulator*.

In general, no channel is ideal. That means that a real channel is subjected

to a large amount of noises and interferences. In order to overcome such noise and interference it is often necessary to increase the transmission power or to introduce some redundancy in the transmitting binary sequence in a controlled manner in order to aid the receiver in decoding the sequence.
In channel where the bandwidth is reduced and the power is available to overcome channel degradations there is no redundancy in the transmitted signal because of the risk of bandwidth consumption due to the added bits. That is why not all the digital communication systems employ a channel encoder.
For what concerns the modulator suppose that the information is transmitted 1 bit per time at a rate B bit/s. If the modulator implements a binary modulation it will map the bit 0 into a waveform $s_1(t)$and the bit 1 into a different waveform $s_2(t)$, allowing each bit from the channel encoder to be transmitted separately. If the modulation is not binary the modulator may transmit K information bit at a time by using $M = 2^k$ distinct waveforms. This implies that a sequence of K bits enter the modulator and, if an equal rate is considered, it takes K times the time period of a binary modulator to obtain the corresponding waveforms.

2.3.1 Error control coding

There are two different kind of codes that are currently used [7]: block codes and convolutional codes (see also paragraph 4.1.1).

The encoder of an (n, k) code for block codes divides the information sequence into k bits message blocks so that the message is represented by a binary k-tuple $m = (m_1, m_2), \ldots, m_k$. It follows that there can be 2^k different messages each of them is transformed in a codeword by the encoder. This means that for each one of the possible 2^K k-bits messages there exists one code word of length n, therefore the code used is called (n, k) *block code.*
One of the main problems in designing the encoder is to choose the number of redundant bits to achieve reliable transmission over noisy channels.
Some useful definitions are following in order to understand the code used for the simulations:

- A block code is *linear* if and only if the modulo-2 sum of two code words is also a code word.

- Given the binary n-tuple $v = (v_1, v_2, \ldots, v_n)$ the *Hamming weight* of v is defined as the number of nonzero components of v.

- Assuming that v and z are two code words the *Hamming distance*, denoted by d, between them is defined as the number of places where they differ and it can be written:

$$d(v, z) = w(v + z). \tag{2.2}$$

- The *minimum distance* of a linear block code is equal to the minimum weight of the linear code:

$$\begin{aligned} d_{min} &= \min\{w(v+z) : v, z \in \mathbf{C}, v \neq z\} = \\ &= \min\{w(x) : x \in \mathbf{C}, x \neq 0\} = \\ &\triangleq w_{min}. \end{aligned} \tag{2.3}$$

- Assuming that p_e is the bit error probability by transmitting a word and noticing that it depends from the ratio E_b/N_0 where E_b is the energy per information bit and N_0 is the (one-sided) noise power spectral density, the *coding gain* (in decibels) is defined as the difference between the E_b/N_0 ratio needed to achieve a bit error probability with coding and without coding.

The encoder for convolutional codes has the same functionalities of the one for block codes but it produces code words that are dependent not only on the corresponding k-bits message block but also on m previous message blocks. That is way the encoder is said to have a m *memory order*. In this case the major problem related to the design of the encoder is how to choose the memory's size in order to achieve reliable transmission over noisy channels.

As the codes that are used in the presented simulations are Reed-Solomon codes, a brief explanation of them and their advantages is following.

Reed-Solomon codes

Reed-Solomon codes are a subset of BCH codes ([7]) and are linear block codes. A Reed-Solomon code is specified as RS(N, K) with m-bit symbols.

Block length	$N = m - 1$
Number of parity check-digits	$(N - K) = 2t$
Minimum distance	$d_{min} = 2t + 1$

Table 2.2: Characteristics of a t-error correcting (N, K)RS code with symbols from $GF(m)$.

This means that the encoder takes K data symbols of m bits each and adds parity symbols to make an N symbol code word.
RS codes achieve the *largest possible* code minimum distance for any linear code with the same encoder input and output block lengths [14]. For non-binary codes the distance between two codewords is defined (analogous to Hamming distance) as the number of symbols in which the sequences differ. For RS codes the minimum distance is given by:

$$d_{min} = (N - K + 1) \tag{2.4}$$

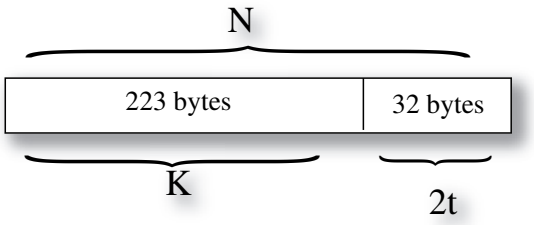
The code is capable of correcting any combination of t or fewer errors, when t can be expressed as:

$$t = \left\lfloor \frac{d_{min} - 1}{2} \right\rfloor = \left\lfloor \frac{N - K}{2} \right\rfloor. \tag{2.5}$$

where $\lfloor x \rfloor$ means the largest integer not to exceed x.
Assuming that an *erasure* is an error with a known location, the erasure-correcting capability, ρ, of the code is:

$$\rho = d_{min} - 1 = N - K. \tag{2.6}$$

It has to be noted that when dealing with nonbinary symbols, each made up of m bits, only a small fraction of possible N-tuples are code words, that means that a large d_{min} can be created with increasing value of m. Another important property of RS codes is that while any other (N, K) linear code is capable of correcting $N - K$ symbol erasure patterns if the erasure symbols all happen to lie in the parity symbols, RS codes are able to correct *any set* of $N - K$ symbol erasures within the block. Table 2.3.1 summarizes the characteristics of a t-error correcting (N, K) RS code, while figure 2.7 shows a popular RS code.



8 bit symbols
N=255
k=223
s=8
2t=32
The decoder can correct any 16 symbol errors in the codeword

Figure 2.7: (255,223)Reed-Solomon code

2.3.2 Modulation for the additive white gaussian noise channel

The description of all possible modulation techniques is a wide topic, that is why all the modulation formats used in the thesis are described with a certain detail in the following sections.

Gaussian minimum shift keying

This modulation is a succession of modifications of binary phase shift keying scheme. For this modulation:

$$s_0(t) = A\cos(\omega t) \qquad \text{represents binary "0"} \tag{2.7}$$

$$s_1(t) = A\cos(\omega t + \pi) \qquad \text{represents binary "1"} \tag{2.8}$$

Quadrature phase shift keying (QPSK) is derived from BPSK defining four signals, each with a phase shift differing by $\pi/2$. The input binary stream $\{d_k\}$, $d_k = 0, 1, 2, \ldots$ is separated at the modulator input into two data streams $d_I(t)$ and $d_Q(t)$ containing even $d_0, d_2, d_4 \ldots$ and odd $d_1, d_3, d_5 \ldots$ bit respectively. A convenient orthogonal modulation of a QPSK modulation is achieved by modulating the in-phase and quadrature data stream into the

cosine and sine function of a carrier wave as follows:

$$s(t) = \frac{1}{\sqrt{2}} d_I(t) \cos(2\pi f t + \pi/4) + d_Q(t) \sin(2\pi f t + \pi/4)$$

using trigonometric identities this can also be written as follows:

$$s(t) = A \cos(2\pi f t + \pi/4 + \theta(t)) \tag{2.9}$$

The pulse stream $d_I(t)$ modulates the cosine function with an amplitude of ± 1. This is equivalent to shift the phase of the cosine by 0 or π; consequently this produces a BPSK waveform. Similarly the pulse stream $d_Q(t)$ modulates the sine function yielding a BPSK waveform orthogonal to the cosine function. The summation of these two orthogonal waveform is the QPSK waveform. Each of the four possible phases of the carrier represents two bits of data. Thus there are two bits per symbol. Since the symbol rate for QPSK is half the bit rate, twice as much data can be carried in the same amount of channel bandwidth as compared to BPSK.
If the two bit streams I and Q are offset by a 1/2 bit interval, then the amplitude fluctuations are minimized since the phase never changes by π. This is a new modulation scheme obtained from QPSK called offset quadrature phase shift keying and it can be demonstrated that the delay of the phase has no effect on the error or bandwidth.
Finally, gaussian minimum shift keying is obtained from OQPSK by replacing the rectangular pulse with a gaussian-shape pulse. This kind of pulse generates a signal with low side lobes and narrower main lobe than the rectangular pulse.

Differential binary phase shift keying

It is also derived from BPSK with the difference that change of bits are modulated instead of bits themselves.

Gaussian frequency shift keying

It derives from the frequency shift keying technique which signals can be represented as follows:

$$s(t) = \sqrt{\frac{2E_s}{T}} \cos\left(2\pi f t + \frac{a_n h(t - nT)}{T} + \theta\right) \qquad nT \leq t \leq (n+1)t \tag{2.10}$$

where E_s is the energy of the signal, T is the symbol duration, f is the carrier frequency, h is the modulation index, a_n is the nth data bit and θ is a constant phase shift. In FSK the signal essentially switches between two frequencies. The modulation index defines how much these two frequencies are spaced and it is defined as the multiplication between the frequency variation that characterizes the two logical values 1 and 0 and the symbol period. The difference between GMSK and GFSK is related to this modulation index that has to be exactly 0.5 for GMSK and is allowed to vary between 0.1 and 1 for FSK.
Gaussian frequency shift keying is simply frequency shift keying but the input is first passed through a gaussian filter.

2.3.3 Introduction to M-ary modulations

For a given M-ary alphabet composed by $0, 1, \ldots, M-1$ symbols, each symbol is related to a unique sequence of k bits expressed as:

$$M = 2^k \qquad \text{or} \qquad k = \log_2 M \tag{2.11}$$

where M is the size of the alphabet.
Each symbol of the M-ary alphabet is mapped onto an electrical voltage or current waveform [3] and is transmitted during each symbol duration T_s.
In a general M-ary signaling system, the possible transmitted signal waveforms are denoted as $\{s_m(t)\}$, $m = 1, 2, \ldots, M$; usually these waveforms are bandpass and, hence, they are represented as [4]:

$$s_m(t) = \mathrm{Re}[u_m(t)e^{j2\pi f_c t}] \qquad m = 1, 2, \ldots, M \tag{2.12}$$

where $u_m(t)$ denote the equivalent low pass waveforms.
M-ary modulations observed in this article are M-ary phase shift keying (MPSK), M-ary frequency shift keying (MFSK), M-ary quadrature amplitude modulation (MQAM). A brief explanation of them is following, [4].

M-ary phase shift keying

The general representation for a set of M-ary phase signaling waveform is:

$$s_m(t) = \mathrm{Re}[u(t)e^{j(2\pi f_c t + \frac{2\pi}{M}(m-1)+\lambda)}] \qquad m = 1, 2, \ldots M \qquad 0 \le t \le T_s \tag{2.13}$$

[3] this is why the terms "waveforms" and "symbols" are sometimes used interchangeably

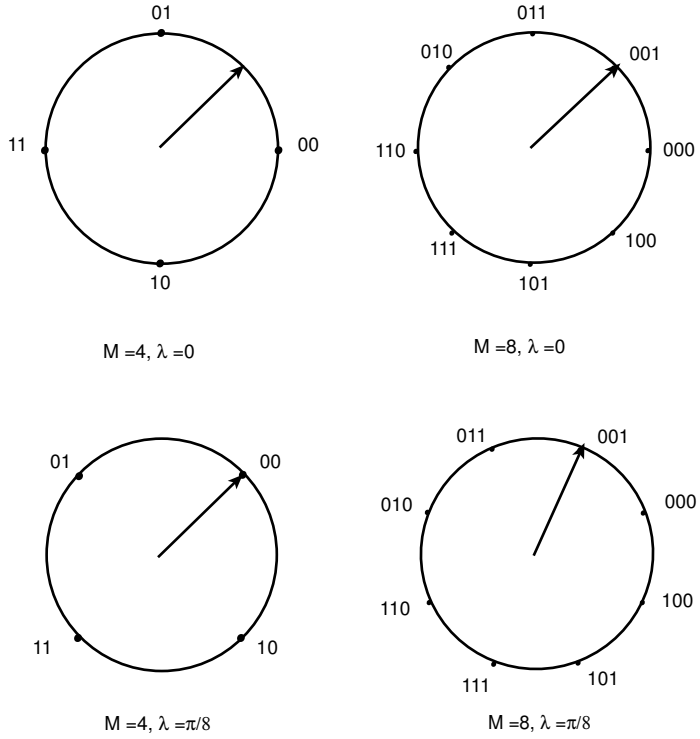


Figure 2.8: PSK signal constellation for $M = 4$ and $M = 8$.

where λ is the initial phase, $u(t)$ is a rectangular pulse that as an amplitude of A if the modulation is an M-PSK (otherwise this would be a general *multiphase signaling technique*). A 4-PSK constellation and a 8-PSK constellation with initial phases equal to 0 and $\pi/8$ are shown in figure 2.8. For what concerns the decision error it is made if the noise causes the phase to fall outside the range $-\pi/M \leq \theta \leq \pi/M$. Thus:

$$P_M = 1 - \int_{-\frac{-\pi}{M}}^{\frac{\pi}{M}} p(\theta)\, d\theta \tag{2.14}$$

An approximation of the error probability for large values of M may be obtained by approximating at first $p(\theta)$. It can be shown that:

$$P_M \approx \operatorname{erfc}\left(\sqrt{\frac{Ay}{N_0 B} \cdot \frac{Kx+1}{1+x}} \cdot \sin\frac{\pi}{M}\right) \tag{2.15}$$

The preferred assignment of k information bits to the $M = 2^k$ possible phases is called *Gray encoding* and it makes the adjacent phases differing by one binary digit as illustrated in figure 2.8. This coding technique is useful because the most likely errors in demodulation occur between adjacent symbols which implies an erroneous selection of an adjacent phase which differ from the correct one only for one binary digit. When a Gray code is used, the bit error probability is also well approximated by:

$$P_b \approx \frac{1}{k} P_M \approx \frac{1}{\log_2 M} P_M \tag{2.16}$$

M-ary frequency shift keying

This is a kind of orthogonal signaling which means a group of signals having a cross correlation coefficient $\rho_r = 0$. The waveforms can be represented as:

$$s_m(t) = \mathrm{Re}\left(A e^{j2\pi(f_c + (2m-3)\frac{\Delta f}{2})t} \right) \qquad m = 1, 2, \ldots M \qquad 0 \leq t \leq T_s \tag{2.17}$$

Where:

$$\Delta f = \frac{m}{2T_s} \tag{2.18}$$

Since the signaling waveforms are orthogonal the decision variables consist in noise only and it can be demonstrated that the probability of a symbol (k bit character) error is upperbounded by:

$$P_M \leq \frac{M-1}{2} \mathrm{erfc}\left(\sqrt{\frac{\eta}{2}} \right) \tag{2.19}$$

where:

$$\eta = \frac{E_b}{N_0} \tag{2.20}$$

The bit error probability is given by:

$$P_b = \frac{M}{2(M-1)} P_M \tag{2.21}$$

M-ary quadrature amplitude modulation

It uses a combination of multiple phases and amplitudes to transmit the k bit information symbols.

The general form for the combined multiple phase and multiple amplitude signal is:

$$\begin{aligned} s_m(t) &= c_m \cdot \cos 2\pi f_c t + \theta_m = \\ &= A_m \cos(2\pi f_c t) + B_m \sin(2\pi f_c t) \\ & \quad m = 1, 2, \ldots M \\ & \quad 0 \leq t \leq T_s \end{aligned} \tag{2.22}$$

where $\{A_m, B_m\}$ is a set of discrete amplitudes. Since these waveforms consist of two phase-quadrature carriers, we can refer to the resulting modulation technique as quadrature amplitude modulation (QAM).
The received signal is demodulated coherently.
Note that if k is even the $M = 2^k$ signal points result in asymmetrical form of QAM which may be view as two separate PAM (see [4] impressed on phase-quadrature carriers, we obtain the following symbol error probability:

$$\begin{aligned} P_M &= 2\left(1 - \frac{1}{\sqrt{M}}\right) \operatorname{erfc}\left(\sqrt{\frac{3}{2(M-1)} k\eta_b}\right) \cdot \\ & \quad \cdot \left(1 - \frac{1}{2}\left(1 - \frac{1}{\sqrt{M}}\right) \operatorname{erfc}\left(\sqrt{\frac{3}{2(M-1)} k\eta_b}\right)\right) \end{aligned} \tag{2.23}$$

When a Gray coding is used to assign k information bits to the $M = 2^k$ possible signals the bit error probability is given by:

$$P_b \approx \frac{1}{\log_2 M} \cdot P_M \tag{2.24}$$

Bandwidth/power limited systems

Since one of the M symbols or waveforms is transmitted during each symbol duration and as each symbol is composed by k bits, the data rate can be expressed as:

$$B = \frac{k}{T_s} = \frac{\log_2 M}{T_s} \qquad \frac{\text{bit}}{s} \tag{2.25}$$

So the effective time duration of each bit would be:

$$T_b = \frac{1}{B} = \frac{T_s}{k} = \frac{1}{kR_s} \tag{2.26}$$

and the symbol rate is:

$$R_s = \frac{B}{\log_2 M} \tag{2.27}$$

From equations 2.25 and 2.26 any digital scheme that transmits $k = \log_2 M$ bits in T_s seconds using a bandwidth of W Hz, operates at a bandwidth efficiency of:

$$\frac{B}{W} = \frac{\log_2 M}{WT_s} = \frac{1}{WT_b} \tag{2.28}$$

This equation implies that the smaller the WT_b product, the more bandwidth efficient would be any digital communication system. Signals with small WT_b product are often used with bandwidth limited systems (i.e. GMSK for GSM), where the object is maximizing the transmitted information rate within the allowable bandwidth at the expense of E_b/N_0 (while maintaining specified values of bit error probability). On the other hand signals with greater WT_b product are used for power limited systems, where the aim is to reduced the required E_b/N_0 at the cost of increased bandwidth. In particular MPSK is used for bandwidth limited systems and it allows a bandwidth efficiency of:

$$\frac{B}{W} = \log_2 M \qquad \frac{\text{bit}}{s}\text{Hz} \tag{2.29}$$

this is because, assuming that $u(t)$, 2.13, is a pulse of duration T_s and that its bandwidth W is approximately equal to the reciprocal of T_s, it follows that $W = 1/T_s$ and sice $T_s = k/B$ than $W = B/\log_2 M$. As M increases in value, also B/W increases, that is why MPSK modulation is a bandwidth efficient scheme and it can be used to improve bandwidth efficiency at the cost of E_b/N_0. Previous considerations are still valid for QAM.
On the other hand noncoherent orthogonal MFSK is power efficient. In fact, as the channel bandwidth required for the transmission is :

$$W = M \cdot \Delta f = \frac{M}{T_s}$$

and the corresponding transmission rate is:

$$B = \frac{k}{T_s} = \frac{1}{T_s}\log_2 M$$

it follows that the bandwidth efficiency is given by:

$$\frac{B}{W} = \frac{\log_2 M}{M} \qquad \frac{\text{bit}}{s}\text{Hz} \tag{2.30}$$

where $\frac{B}{W}$ decreases as M increases. So MFSK modulation can be used for realizing a reduction in required E_b/N_0 at the cost of increased bandwidth. When error correcting code is considered, modulation selection is not so simple because coding techniques can provide power bandwidth tradeoffs more effective than would be possible through the use of any modulation technique.

Chapter 3

Problems of TCP over wireless

There are two main problems that TCP communications over wireless channels must address. The first problem refers to the high bit error rate (BER) that a wireless channel has. High BER causes corruption in the data transmitted over the link which may cause the loss of TCP data segments or acknowledgments.

The second problem is related to the fact that this interruption in the flow of communication is interpreted by TCP as a congestion situation. That is why TCP incorrectly initiates its congestion control algorithm or congestion avoidance mechanism if one loss is notified. Mobile host to mobile host communications involves at least two wireless links therefore the effect of high BER may cause even more segments and acknowledgments to be lost making it more likely for TCP to incorrectly assume congestion on the link.

Moreover the TCP sawtooth behavior itself, due to the congestion control algorithm, causes delays and lower response time because of the queues felt to the limit.

Chapter 4

Previously proposed solutions

A lot of solutions have been proposed in literature to solve the problems related to TCP over wireless that have been summarized in the previous section. A short introduction to the most important is following:

1. *Split connection approach* considers the TCP connection as terminated at the base station or at the access point between the wired and the wireless link, while adopting a different transmission protocol for the wireless connection. This implies that different flow error control protocols, packet size and time outs would be used for each part.
   There are of course some drawbacks:

   - TCP end-to-end semantics are violated because the sender can get the acknowledgement before the receiver gets the packet.
   - base stations failures in sending acknowledgements to the sender may result in data losses as shown in figure 4.2.
   - the handoff is much more complicated due to state information at the access point or base station where the protocol is split.

2. *End-to-End approach* consists in an explicit notification of packets lost in the wireless link. For instance, if the sender receive an acknowledgement marked with an Explicit Loss Notification, it will perform retransmission without invoking the associated congestion control procedures. This solution maintains the end-to-end semantics of TCP and do not require extra overhead at the base station for protocol processing or handoff. The drawbacks are:

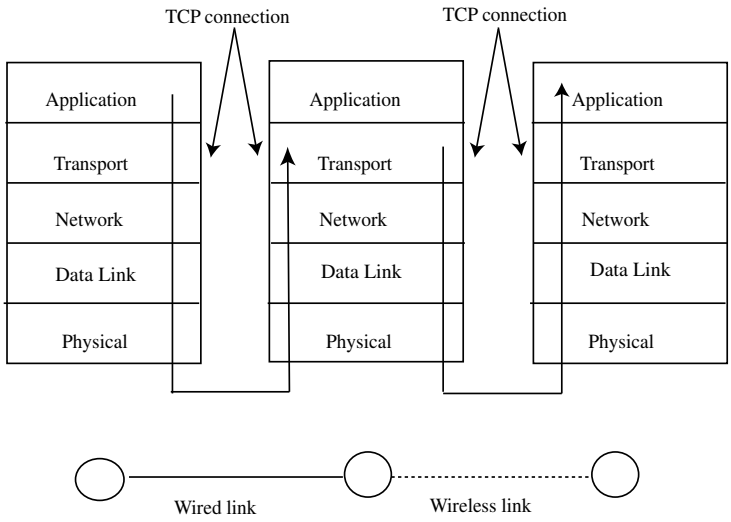


Figure 4.1: Split connection approach, TCP end-to-end semantics are not preserved.

- the link layer at the base station has to be TCP aware.
- it cannot be used if TCP data and TCP ACKs traverse different paths.
- it requires additional space to the base station for extra information.

3. *Link layer approach:* it tries to reduce radio link errors through radio resource management techniques like Forward Error Correction of damaged packets, Automatic Repeat Request of damaged packets, Power Control and Power Allocation, Rate Adaptation etc. Of course each solution presents some drawbacks. For example, considering the forward error correction, it adds to the original data some redundant information improving the quality of the noisy link but, on the other end it consumes some extra bandwidth and requires some processing time for coding and decoding the redundant information. In general this solution requires no change to the existing sender behavior and matches layered protocol model but it will cause interactions with TCP, e.g., fast retransmission by TCP can be triggered by delays due to link-level timeout and retransmission.

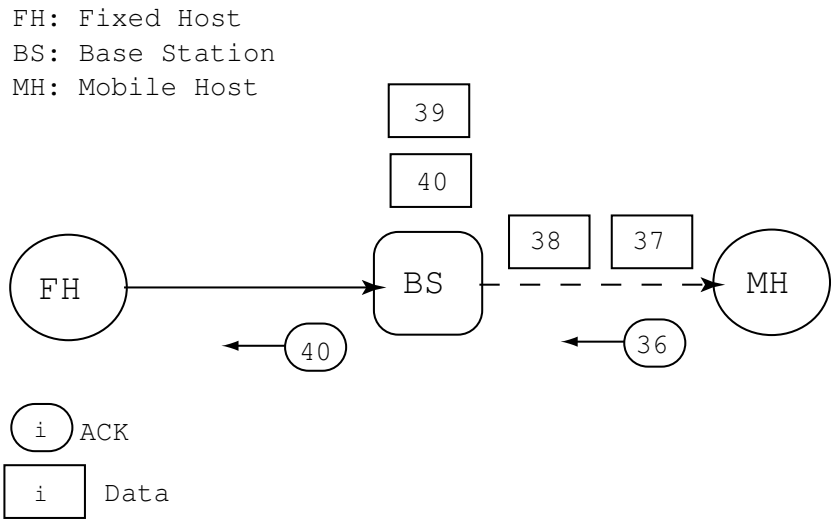


Figure 4.2: Split Connection drawback: the sender does not buffer 40 after receiving its ACK, but, if the Base Station fails in sending to the receiver 40, this will be lost.

The aim of this work is to focus on the link layer approach, in particular using both FEC and ARQ to solve the problem that TCP interprets all the losses as due to congestion. The effect of the amount of the bandwidth consumed by both FEC redundancy and different modulations used is also considered in order to get the maximum gain in TCP performance.

4.1 Techniques to improve TCP reliability: Link Layer approach

As already said in chapter 1, adaptation for limited and time varying wireless links can be reached at different layers and it has to be observed that also the following techniques, presented for the link layer, can be implemented in the upper layers.

4.1.1 FEC: Forward Error Correction

Forward error-correction coding (also called channel coding) is a type of digital signal processing in which a transmitter of digital data adds extra in-

formation, known as redundant bits, to the data stream. This check sequence enables a receiving system to detect and possibly correct errors caused by corruption from the channel and the receiver. As the name implies, this coding technique enables the decoder to correct errors without asking for retransmission of the original information. In a communication system that employs FEC, a digital information source send a data sequence to an *encoder*. This encoder inserts redundant (or *parity*) bits, thereby outputting a longer sequence of code bits, called a codeword. Such codewords can then be transmitted to a receiver, which uses a suitable *decoder* to extract the original data sequence. Codes are designated with the notation (n, k) according to the number of n output code bits and k input data bits. The ratio $R_c = k/n$ is called the *rate of the code* and is a measure of the fraction of information contained in each codeword or as the number of the information bits sent per transmitted symbol.
The exact methods used for detection and correction depend on the type of information being transmitted and the form in which it was transmitted.

The basics of these methods were laid down in 1948-49 by Claude Shannon in the *Mathematical Theory of Communication*, where he purposed the algebraic coding technique (also known as *block coding*). With this technique the encoder intersperses parity bits into the data sequence using a particular algebraic algorithm and the decoder applies an inverse of this algorithm to identify and correct error caused by channel corruption.

Another forward-error correcting technique, known as *convolutional coding*, was first introduced in 1995. The paramount feature of such codes is that the encoding of any bit is strongly influenced by the bits that preceded it (that is, the memory of past bits). A convolutional decoder takes into account such memory when trying to estimate the most likely sequence of data that produced the received sequence of code bits. Such procedures require great deal of memory and typically suffer from buffer overflow and degradation.

In 1967 Andrew Viterbi developed a decoding technique in which at each bit-interval, the Viterbi decoding algorithm compares the actual received code bits with the code bits that might have been generated for each possible memory-state transition. This technique requires less memory than the one is needed for convolutional decoding.

In 1993 Claude Berrou and his associates developed the *turbo code*, [25], the most powerful forward error correction code yet. Using the turbo code, communication systems can approach the theoretical limit of channel capacity.

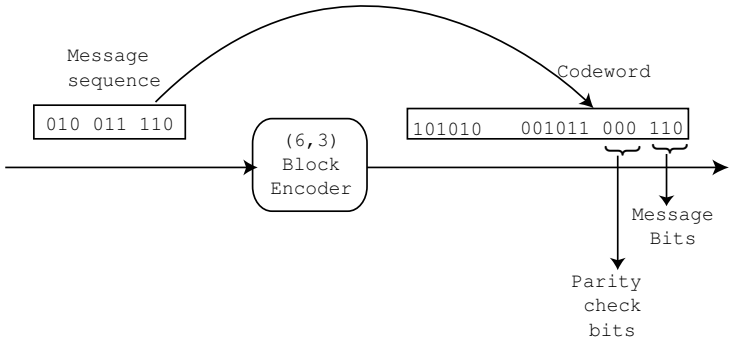


Figure 4.3: An example of a $(6, 3)$ algebraic encoder that produces a six-bit codeword for every three-bit message sequence. In this example, each six-bit output codeword is composed of the original three-bit message sequence and a three-bit parity sequence. This codeword format is known as *systematic*.

In general the greater the number of check bits added, the greater the error-correcting capability of the given code. Of course, after a time, the check bits themselves become the source of errors and the check bits use up bandwidth that could otherwise be used for data bits. The efficiency of a coding scheme is measured by the number of redundant bits that must be added to detect and correct a given number of errors.
FEC systems are classified by the number of bits added to the data stream. For example, in some FEC systems, the number of redundant bits is equal to the number of data bits, this makes the total bit rate double that the original data. This is known as a "rate one half" scheme.

4.1.2 ARQ: Automatic Repeat Request

Automatic repeat request in an error-control system in which a request for retransmission is generated by the receiver when an error in transmission is detected. A certain number of systems can be used. In a simple system a positive acknowledgement (ACK) is returned when the data is received correctly and a negative acknowledgement (NACK) is returned when an error is detected. There are three main types of ARQ:

1. *Stop and Wait*: in this solution the sender transmits a block of data and then waits for an acknowledgement before transmitting the next block.

2. *Selective repeat*: in this case the sender only retransmits blocks that are in error and it does not wait for an ACK before transmitting the next block.

3. *Go back N*: also in this case the sender does not wait the previous ACK for the transmission of the next block and retransmission is done for all the blocks sent after the one to which the NACK is referred

Note that the loss of a block can also be notified at the transport level by the use of triple-duplicate ACKs (which refer to a block previously sent) and the expiration of a Time Out.

4.1.3 Hybrid ARQ schemes

The drawback of ARQ is that its throughput falls rapidly when the channel error rate is high, like in a wireless channel, due to the increased frequency of retransmissions. On the other hand, more powerful FEC error correction capabilities, imply more redundancy, more expensive hardware and longer potential delays.
A solution to the problems caused by the previous error correcting methods, could be represented by the *Hybrid Automatic Repeat Request* which consists in a FEC subsystem contained in an ARQ system. FEC has to correct the error patterns that occurs most frequently in order to reduce the frequency of retransmission, ARQ is used for retransmission when a less frequent error pattern occurs and is detected. In this way it is possible to achieve a higher reliability than a FEC system alone and a higher throughput than a system with only ARQ. There are two types of Hybrid ARQ:

- *Type I Hybrid ARQ scheme*- It uses a code, i.e. an (n, k) linear code, which is designed for simultaneous error detection and error correction. When the received block is detected with errors, the receiver first attempts to locate and correct the errors. If the number of errors is within the error-correcting capability of the code the errors will be corrected and the decoded message will be passed to the user or saved in

a buffer until it is ready to be delivered. If an uncorrectable error pattern is detected, the receiver rejects the received block and requests a retransmission. This error-correction and retransmission continues for a fixed number of times called *persistency* or, in its absence, until the block is correctly received. Since a code used in this scheme must be able to correct a certain collection of error patterns a to detect other error patterns, more parity check digits are needed. This increases the overhead for each transmission and retransmission and when the channel error rate is low, it has a lower throughput than its corresponding ARQ scheme. However when the channel error rate increases, the ARQ schemes throughout drops rapidly and the hybrid ARQ scheme provides a higher throughput.

- *Type II Hybrid ARQ scheme*-It is based on the concept that parity check digits are sent to the receiver only when they are needed. Two linear codes are used in this type of scheme: one is a high rate code which is designed for error detection only, the other is an invertible half-rate code which is designed for simultaneous error correction and error detection [7].

4.1.4 Power Control

The limited bandwidth of radio frequencies imposes severe restrictions on the planning of radio networks. In order to produce high quality and high capacity communications services, efficient methods of radio resource management are needed for sharing this narrow band among several users [10]. In practice, sharing the bandwidth always gives interference that is directly proportional to the transmission power. Controlling the transmission power is one of the key techniques for balancing the received signal and the interference.
In 3rd genaration mobile communication systems where the Code Division Multiple Access technique is used to allow many users to simultaneously access a given frequency allocation, the not perfect orthogonality of the spreading sequences used can cause the so called Multiple Access Interference. Unintended transmissions add nonzero MAI during the dispreading at a receiver. A severe instance of MAI is the so called *near-far problem* and it happens whereby a receiver who is trying to detect the signal of the i-th transmitter may be much closer in distance to, say, the j-th transmitter than the i-th

transmitter. When all transmission powers are equal, the signal from the j-th transmitter will arrive at the receiver in question with a substantially larger power than that of the i-th transmitter, causing incorrect decoding of the i-th transmission (i.e., a secondary collision). In order to minimize the near-far problem, the goal in CDMA systems is to assure that all mobiles achieve the same received power levels at the Base Station. The target value for the receive power level shall be the minimum level possible which allows the link to meet user-defined performance objectives (Bit Error Rates, Frame Error Rates, Capacity Estimates, Dropped-call Rate Estimates, and Coverage goals). In order to implement such a strategy, mobiles that are closer to the Base Station must transmit less power than mobiles that are further away from the Base Station. For 3rd generation wireless systems two power control mechanism are commonly used: a fast inner loop power control and a slower outer loop power control [15]. This two mechanisms are actually implemented in the physical layer but they are strictly related to the solutions purposed in this section that is why they are cited here.

4.2 Objective of the thesis

This thesis is focused on the link layer approaches previously described. In particular, basing on [1], the aim is to achieve optimal TCP performances through the use of:

- Hybrid-ARQ I type:
  - FEC is implemented through the use of RS codes.
  - ARQ is based on the Selective Repeat technique, this means that, when the sender receives a negative acknowledgement (NACK), the packet to be retransmitted has the maximum priority.
- Different modulation formats.
- Different values of the transmission power

Chapter 5

Model for TCP over a Wireless Link

Considering only one sender, only one receiver and only one base station for acknowledgements crossing, the model obtained for TCP extended over wireless link is shown in figure 5.1, [1]. It has been assumed that TCP packets of size MSS (expressed in bits), traversing both wired and wireless links, are subdivided in blocks for the transmission in the wireless part. This implies that one packet in the wired part becomes equal to J blocks:

$$J = \frac{MSS}{K} \tag{5.1}$$

where, obviously, K is the length of each block. Each block is then encoded and transmitted as shown in figure 2.6.

Furthermore it has been assumed that the base station has large enough buffer and that acknowledgements for a packet traverse through the same base station of the packets.
An hybrid-ARQ of the first type has been used in order to improve TCP performances. Due to the use of a FEC encoding, each K bits-block is encoded in a code word of length N.
The definition of the redundancy ratio follows:

$$x = \frac{N - K}{K} \tag{5.2}$$

We assume that a strong CRC code is used so that the probability of not detecting a corrupted block is zero. We also assume that ACK are well

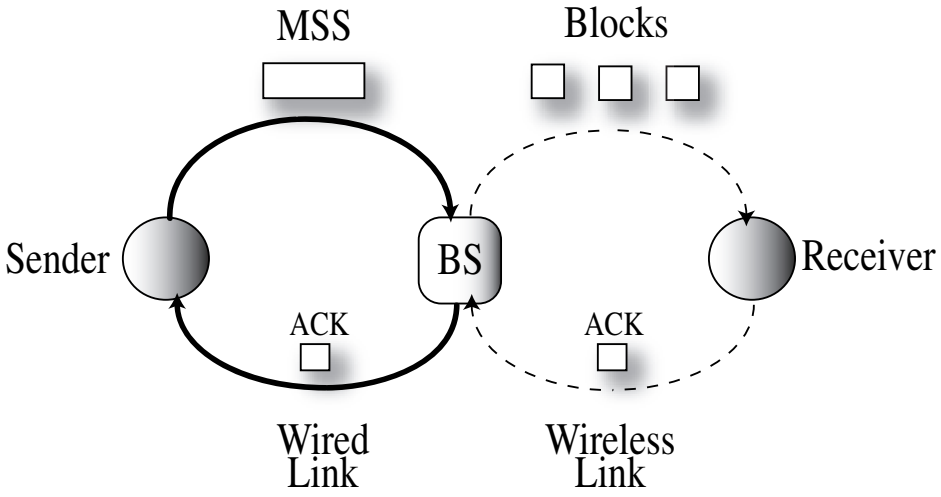


Figure 5.1: A model for a hybrid wired/wireless network

protected and there is no probability to lose them in any link.
A TCP segment is decoded properly if all blocks are received uncorrupted and decoded by the receiver.
In general, for an (N, K) code, the coding gain, which is the reduction in the transmission power achieved using a coded signal, is given by $G_{coding} = \frac{K}{N} d_{min}$. If Reed-Solomon encoding is used (2.3.1):

$$d_{min} = (N - K + 1) \tag{5.3}$$

from equation 5.2 we can write the following:

$$\frac{N}{K} = x + 1 \tag{5.4}$$

it follows that in this particular case the expression of the coding gain can be expressed as:

$$G_{coding} = \frac{1}{x+1}(Kx + 1) \tag{5.5}$$

If FEC does not succeed in decoding one block, the link level error mechanism turns to ARQ and tries to send the frame for δ times where δ is the maximum number of allowable retransmissions called *persistency*, 4.1.3. If the frame cannot go through the link after δ trials ARQ assumes that the frame cannot be locally recovered and leaves to TCP the correction of the

frame on an end-to-end basis.
Moreover we assume that when a loss is notified at the input of the wireless link , the corresponding block is directly retransmitted giving it the maximum priority.

For what concerns the power management it has to be said that, considering a general expression of the error probability, p_e decreases when the ratio $\frac{E_b}{N_0}$ decreases, this means that the link reliability can be improved by increasing the transmission power.[1] Increasing the transmission power improves TCP performances but causes greater energy consumption and aggravates the interferences with other neighboring communications. Note that the relation between energy per bit and transmission power is given by: $E_b = \frac{A \cdot y}{B}$ where y is the transmission power, B is the transmission rate and A is the attenuation of the channel.

[1] Of course the relation between p_e and $\frac{E_b}{N_0}$ depends on the particular modulation technique used.

Chapter 6

Definition and Evaluation of the Objective Function

The aim of this work is to find the link-layer techniques that can improve the values achieved by the Objective Function.
The *Objective Function*, γ, is defined as the ratio between the throughput $\lambda(x, y, \delta)$ and the cost c, i.e.,$\gamma = \lambda/c$, [1]. Basing on the Reno flavor of TCP, which is the most popular implementation of the Internet today, [18], [19], evaluation of both throughput and cost function is made in the following paragraphs [2].

6.1 TCP Throughput Evaluation

For what concerns the expression of the throughput as a function of loss rate and round trip time, not only the behavior of TCP fast retransmit mechanism has been considered but also the effect of TCP timeout mechanism. This is really important as it has been observed that the number of TCP timeouts is usually bigger than the fast retransmit events, [17], [29], [28]. The effects of small receiver-side windows are also explicitly modelled.

TCP congestion avoidance behavior is modelled in terms of rounds.
In particular a round starts with the back to back transmission of W packets, where W is the current size of the congestion window, and finishes when the sender receives the first ACK for one of these packets. In the model it has been assumed that the round duration is equal to the round trip time and

independent on W and that the time needed to send all the packets in a window is smaller than the round trip time. Let 1 be the number of packets that are acknowledged by a received ACK; if W packets are transmitted in the first round than W ACKs are received and since each acknowledgement increases the window size of $1/W$, the window size at the beginning of the second round is $W' = W + 1$.
In the following subsection the behavior of TCP in presence of packet loss is presented. Packet loss can be notified in two ways: by the reception of "triple-duplicate" acknowledgements or by the expiration of a Time Out (TO). It is assumed that a packet is lost independently of any packets lost in other rounds and that if a packet is lost all the subsequent packets in the same round are lost too.

Losses notified only by "triple-duplicate" ACKs

Let N_t to be the number of packets transmitted in the interval [0,t] and f_t=N_t/t the throughput on that interval.
The long term steady state TCP throughput f is defined as follows:

$$f = \lim_{t \to +\infty} f_t \tag{6.1}$$

Defining p as the probability that a packet is lost, given that either it is the first packet in its round or the preceding packets in its round are not lost. The aim is to obtain a relationship $f(p)$ between the throughput of the TCP connection and p, the loss probability defined above.
It can be demonstrated that under this assumption the following equation can be obtained, [2]:

$$f(p) = \frac{1}{RTT}\sqrt{\frac{3}{2p}} + o\left(\frac{1}{\sqrt{p}}\right) \tag{6.2}$$

Losses notified by "triple-duplicate" ACKs and TOs

A TO for a loss occurs when packets are lost and less than tree duplicated ACKs are received.
After a TO the congestion window is reduced to one, and one packet is thus resent in the first round after a TO. If another TO occurs without successfully retransmission of the packets lost during the first TO, the period of timeout doubles of $2T_0$. The doubling is repeated for each unsuccessful

retransmission until $64T_0$. Under this assumptions the following expression for the throughput can be demonstrated [2]:

$$f(p) \approx \frac{1}{RTT\sqrt{\frac{3}{2p}} + T_0 \min(1, 3\sqrt{\frac{3p}{8}})p(1+p^{32})} \tag{6.3}$$

Losses notified by "triple-duplicate" ACKs and TOs with window limitation

It is assumed now that at the beginning of each TCP flow the receiver advertises a maximum buffer size which determines a maximum congestion window size, W_{max}. As a consequence during a period without loss indications, the window size can grow up to W_{max}, but will not grow further beyond this value, figure 6.1. The following assumptions are done. Denoting as W_u the unconstrained window size whose mean is given by $E[W_u]$, then it is assumed that $E[W] \approx E[W_u]$ if $E[W_u] \leq W_{max}$.
Otherwise the following approximation is used: $E[W] \approx W_{max}$ if $W_{max} \leq E[W_u]$.
Under these assumptions the correct expression for the throughput is finally obtained:

$$f(p) \approx \min\left(\frac{W_{max}}{RTT}, \frac{1}{RTT\sqrt{\frac{3}{2p}} + T_0 \min\left(1, 3\sqrt{\frac{3p}{8}}\right) p(1+p^{32})}\right) \tag{6.4}$$

The analytical formulas for TCP throughput can be written as follow:

$$\lambda(x, y, \delta) = \frac{1}{x+1} \cdot f(RTT, P_{loss}) \tag{6.5}$$

Defining P_{loss} as the probability that a TCP segment is discarded because of link errors in the wireless channel, it can be evaluated as follow:

$$P_{loss} = 1 - (1 - P_{block})^J \tag{6.6}$$

where P_{block} is the the block error probability of the input decoder and, if Reed-Solomon codes are used, P_{block} can be approximated as:

$$P_{block} \approx (1 - (1 - p_e^K)^{\delta+1}) \tag{6.7}$$

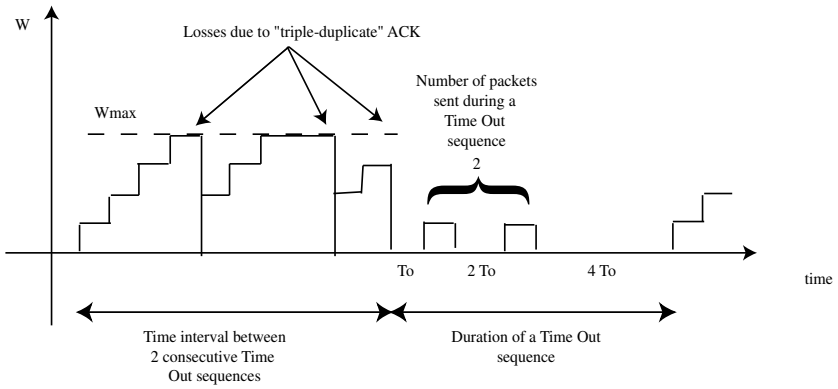


Figure 6.1: Evolution of the window size when limited by W_{max}

As p_e depends on the particular modulation technique used, numerical evaluation of throughput are done using at first three modulations called Gaussian Minimum Shift Keying (GMSK), Differential Binary Phase Shift Keying (DBPSK), Gaussian Frequency Shift Keying (GFSK) and then analyzing the effects of others M-ary modulation techniques.

6.2 Cost Evaluation

Two cost terms are considered, [1]: a term which takes energy consumption into account and a term which considers the amount of wireless resources employed. If x is the redundancy ratio introduced by FEC and if the aim is to transfer S segments each of MSS bits, each segment would be transmitted in a total X of $K(1+x)$ sized codewords over the wireless link, where each codeword could be retransmitted $0 \leq \delta_i \leq \delta$ times.

It follows that $S \cdot MSS \cdot (1+x) \cdot (1+\delta_i)$ bits must be transmitted. Accordingly, the cost of the resources required to complete the transfer is given by:

$$c_{resources} = K_{resources} \cdot S \cdot MSS \cdot (1+x) \cdot (1+E(\delta_i)) \tag{6.8}$$

where $K_{resources}$ (expressed in bit^{-1}) is a constant which represents the cost of the bandwidth resources required to transfer a bit. For what concerns the energy, let call K_{energy}, expressed in $Joule^{-1}$, the cost of a unit of energy. Considering that y/B is the energy transmitted per bit, the total energy cost

can be expressed as:

$$c_{energy} = K_{energy} \cdot S \cdot MSS \cdot (1 + x) \cdot (1 + E(\delta_i)) \cdot y/B \tag{6.9}$$

It follows that the cost of the transfer of S TCP segments can be evaluated as:

$$\begin{aligned} c(x, y) &= c_{energy} + c_{resources} = \\ &= S \cdot MSS \cdot (1 + x) \cdot (K_{energy} \cdot y/B + K_{resources}) \end{aligned} \tag{6.10}$$

6.3 RTT Computation

Assuming that data blocks are quickly acknowledged by ARQ-SR and sizes of acknowledgments are negligible compared to data blocks, the transmission of a block and the reception of its acnowledgment takes $\tau = 2D + \frac{N}{B}$, where$D$ is the one way propagation delay. Let δ_i, $i = 1 \ldots X$, be the number of times we retransmit the block i of a TCP packet. Assuming that the next block is transmitted after the sender receives the acknowledgement of the previous block (i.e. stop and wait), the round trip time of a TCP packet can be written as:

$$RTT = T + 2D + \frac{JN}{B} + \sum_{i=0}^{J} \delta_i \tau \tag{6.11}$$

where T is the round trip time of the wired part of the TCP connection. RTT is a random variable because of δ_i which are i.i.d. and geometric.

$$P_{\delta_i} = \begin{cases} \frac{P_{Block}^k (1 - P_{Block})}{1 - P_{Block}^{\delta+1}} & 0 \leq K \leq \delta, \\ 0 & \text{otherwise} \end{cases}$$

The expected value of the random variable δ_i is computed as:

$$E[\delta_i] = \sum_{j=0}^{\delta} j \frac{P_{Block}^j (1 - P_{Block})}{1 - P_{Block}^{\delta+1}} =$$

$$= \frac{P_{Block}(1 - P_{Block})}{1 - P_{Block}^{\delta+1}} \sum_{j=0}^{\delta} j P_{Block}^{j-1} =$$

$$= \frac{P_{Block}}{1 - P_{Block}} - \frac{(\delta + 1) P_{Block}^{\delta+1}}{1 - P_{Block}^{\delta+1}} \tag{6.12}$$

Therefore the expected RTT, $E[RTT]$ is given by:

$$E[RTT] = T + 2D + \frac{JN}{B} + J\tau P_{Block} \left[\frac{P_{Block}}{1 - P_{Block}} - \frac{(\delta + 1) P_{Block}^{\delta+1}}{1 - P_{Block}^{\delta+1}} \right] \tag{6.13}$$

Chapter 7

Simulations and results

7.1 Effects of the use of Hybrid-ARQ on the Objective Function

In this section the analytical framework developed in the previous section is applied to the three relevant examples of numerical modulation techniques formerly presented:

1. GMSK used in General Packet Radio Service (GPRS).
2. DBPSK used in IEEE 802.11.
3. GFSK used in Bluetooth.

where A is the channel attenuation that is assumed to be caused only by free space losses in the following simulations.
In figure 7.1 the objective function for a GMSK modulation assuming $\delta = 0$ is shown. It has to be observed that the objective function increases with the increasing power level and, for a given value of persistency δ and power y, it first increases to a maximum value and then reduces. This reduction is due to the increasing redundancy that decreases the effective bandwidth. It is clear in fact that the addition of FEC at the link level reduces the loss probability of TCP packets and thus increases the throughput. This improvement continues until the two terms of equation 6.4 become equal. At this point the quantity of FEC added to the wireless link is sufficient to eliminate the negative effect of non-congestion losses on TCP. It is possible to say here that the FEC has *cleaned* the link from TCP point of view, [20]. Any increase of

modulation	error probability
GMSK	$\frac{1}{2}\text{erfc}\left(\sqrt{\frac{Ay}{N_0B}\cdot\frac{Kx+1}{1+x}}\right)$
DBPSK	$\frac{1}{2}\exp\left(-\frac{Ay}{N_0B}\cdot\frac{Kx+1}{1+x}\right)$
GFSK	$\frac{1}{2}\exp\left(-\frac{1}{2}\frac{Ay}{N_0B}\cdot\frac{Kx+1}{1+x}\right)$

Table 7.1: Error probabilities for binary modulations.

redundancy x from this point of view results in a throughput deterioration: more FEC than what is needed to clean the link.
We observe similar behavior in figure 7.2 where we can observe that an increasing of δ for a given power level and redundancy x increases the objective function.
Similiar results have been found for DPSK and GFSK as we can see in figures 7.3 and 7.4

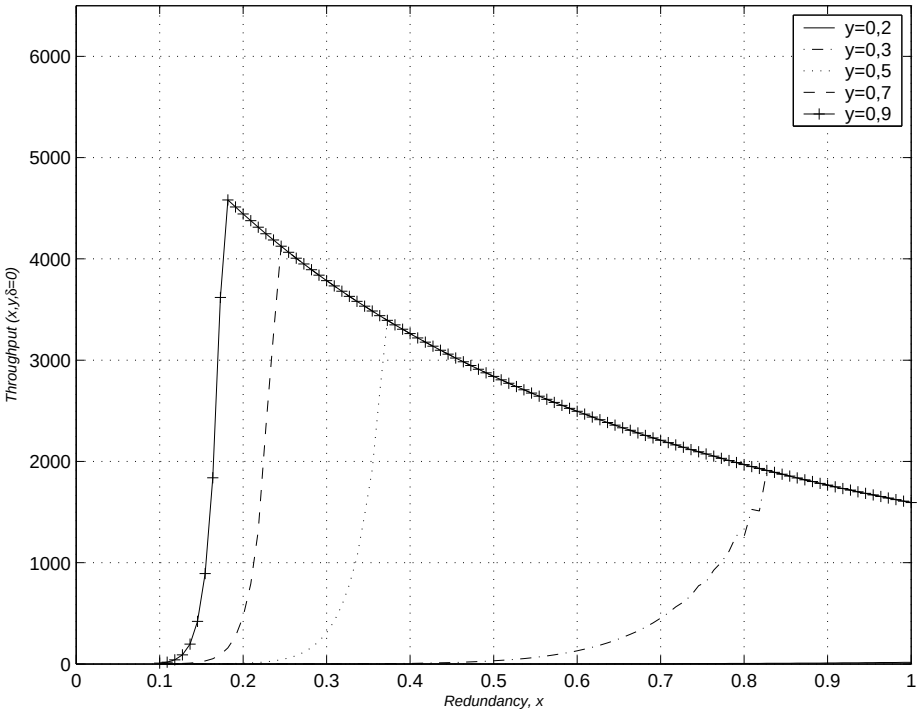


Figure 7.1: GMSK objective function for $\delta = 0$.

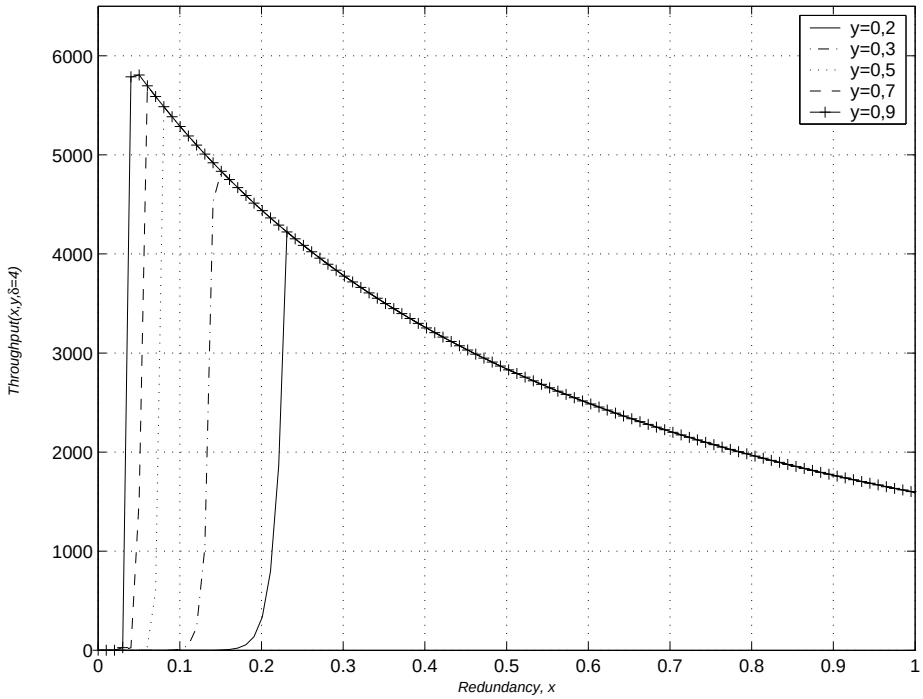


Figure 7.2: GMSK objective function for $\delta = 4$.

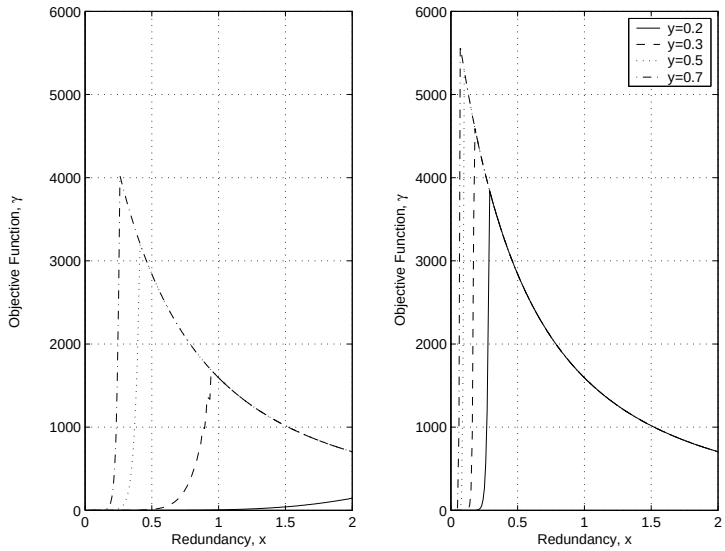


Figure 7.3: DBPSK objective function for δ=0 (graph on the left) and $\delta = 4$ (graph on the right).

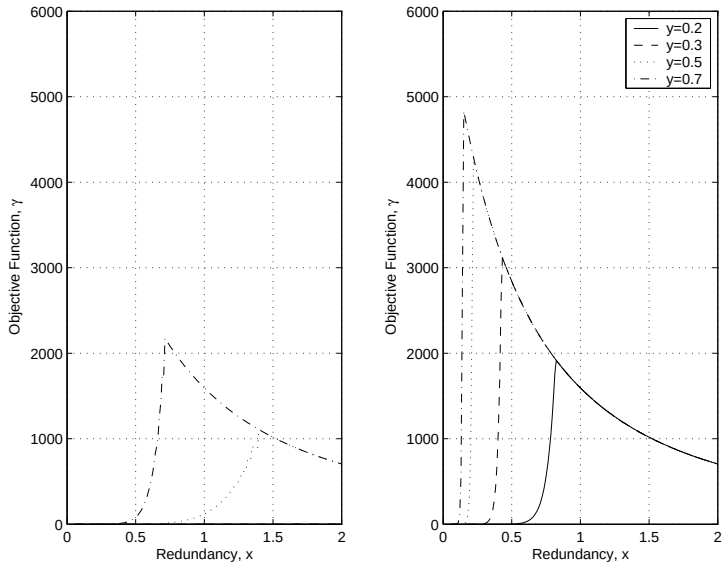


Figure 7.4: GFSK objective function for both $\delta = 0$ and $\delta = 4$.

7.2 Objective Function for M-ary modulation formats

The aim of this section is to compute the value of the objective function for the modulations described in 2.3.3.

7.2.1 Case of absence of bandwidth constraints

If the simulation is done under the assumption that the available bandwidth is infinite than the results obtained are shown in figure 7.5, where the value of the objective function is dependent on the number of symbols M and the redundancy x..
It seems that two different behaviors are obtained by increasing the number of symbols used for the M-ary modulations:

- For MQAM and MPSK modulations the optimal value of the objective function is obtained for $M = 2$ as the performances of the optimization function decrease for any increase of the number of symbols. The behavior of the objective function depending on the redundancy is almost equal to the case of the binary modulation.
- On the other hand, for MFSK modulation, the objective function increases as M increase.

What has been obtained can be explained with the behavior of symbol error probabilities for each modulation. Figure 7.6 and figure 7.7 illustrate these error probabilities as functions of SNR per bit for $M = 2, 4, 8, 16, 32$. The graphs clearly illustrate the penalty in SNR per bit as M is increased beyond $M = 4$ for both MPSK and MQAM.

Moreover, figure 7.8 shows the comparison in performance between MPSK and MQAM. It is clear that there is an advantage in using MQAM instead of MPSK for $M > 4$. These observations explain also the behavior of TCP throughput for the two bandwidth efficient modulations in figure 7.5 where the objective function for a high value of M is better for MQAM than for MPSK. For what concerns the behavior of MFSKs objective function it can also be explained by using the symbol error probability of MFSK shown in figure 7.9. These curves illustrate the advantage of increasing the number of waveforms. That is, by increasing M one can reduce the SNR per bit

Figure 7.5: M-ary modulations objective functions

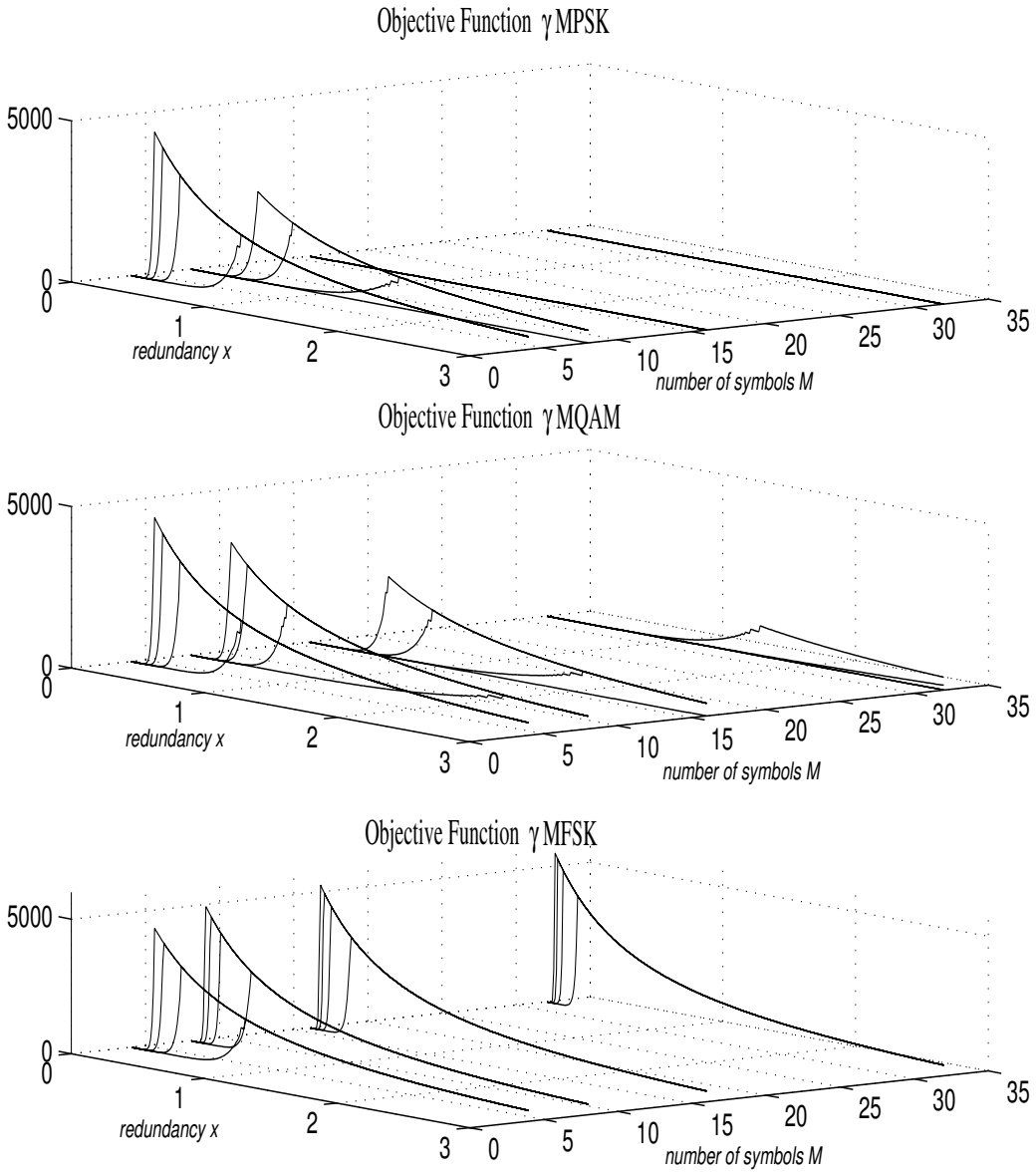


Figure 7.6: Probability of a symbol error for MPSK

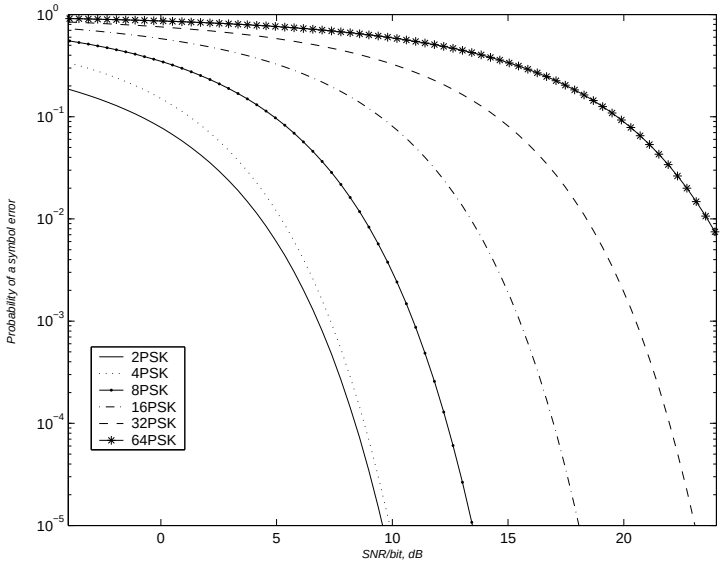


Figure 7.7: Probability of a symbol error for MQAM

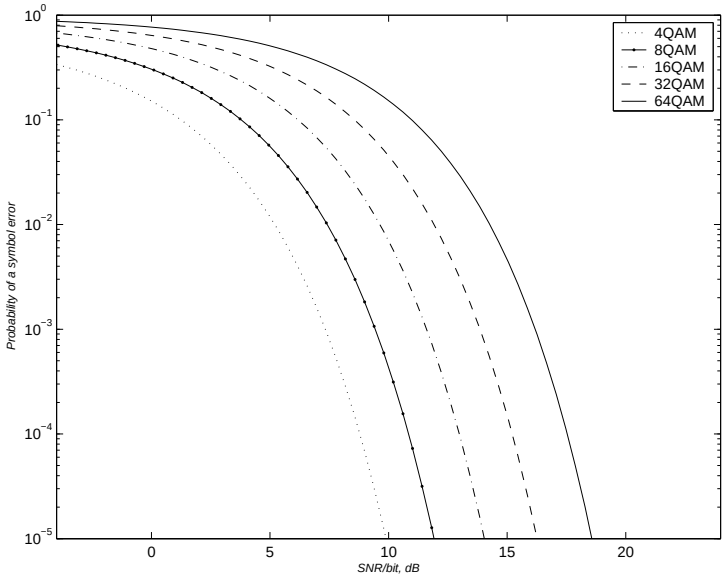


Figure 7.8: Probability of a symbol error for MQAM and MPSK

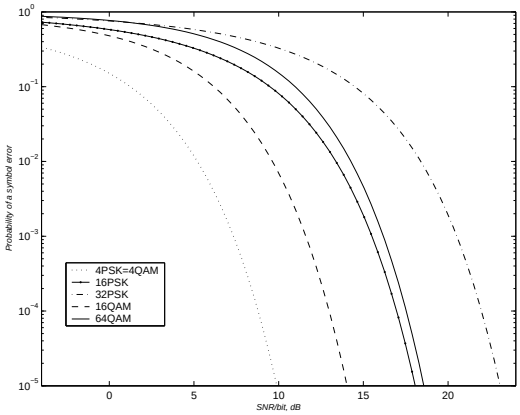
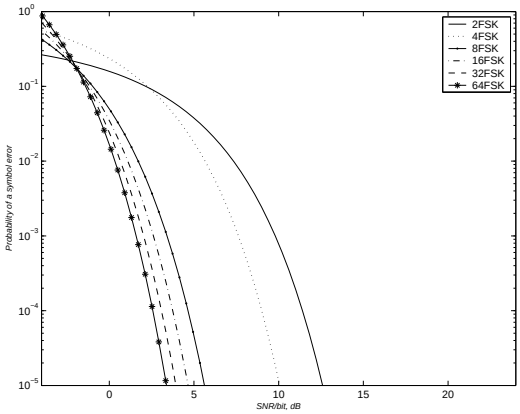


Figure 7.9: Probability of a symbol error for MFSK



required to achieve a given probability of error and so reach greater values of the objective function.

7.2.2 Case of bandwidth constraints

Study cases with bandwidth limitations are now considered. Basing on formulas in section 2.3.3, the problems of objective function maximization are modified as follows:

$$\text{for MPSK and MQAM} \begin{cases} \max\{\gamma(x, y, \delta)\} \\ \\ \frac{x+1}{\log_2 M} \cdot B \leq F \end{cases} \tag{7.1}$$

$$\text{for MFSK} \begin{cases} \max\{\gamma(x, y, \delta)\} \\ \\ \frac{x+1}{\log_2 M} \cdot B \cdot M \leq F \end{cases} \tag{7.2}$$

where F is the available bandwidth and the conditions on the bandwidth are obtained considering expressions 2.29 for MPSK and 2.30 for MFSK. In fact from 2.29 it can be written that:

$$W = \frac{M \cdot B}{\log_2 M} \tag{7.3}$$

As W is the bandwidth required, it has to be imposed that it is not overcoming the available bandwidth.
Similarly it is possible to demonstrate 7.2 starting from 2.30.

Results obtained are shown in figure 7.11 for MQAM and 7.10 for MFSK. Simulations are implemented using the same bandwidth for both the techniques. Note that MPSK format has not been considered because it presents the same behavior of MQAM for the point of view of this study.
Considering the objective function for MQAM is has to be noticed that:

- An increase in the transmission power causes an increase in the objective function and the same level of objective function can be obtained with less redundancy using more power.
- Assuming the use of a high value of the number of symbols M, if the transmission power is not high enough, the available bandwidth is

overcame and M has to be scaled to a higher value in order to maintain a high value of the objective function.

For what concerns MFSK format the following observations can be done:

- A higher objective function can be achieved with a higher value of M, on the other hand, if M is to high the bandwidth will be overcame and the number of symbols has to be decreased.
- Increasing the transmission power, an improvement in the throughput is obtained for a fixed value of M with less redundancy, this implies that scaling to a lower value of M increases also the redundancy for a fixed value of the power.

Figure 7.10: MFSK with bandwidth restrictions

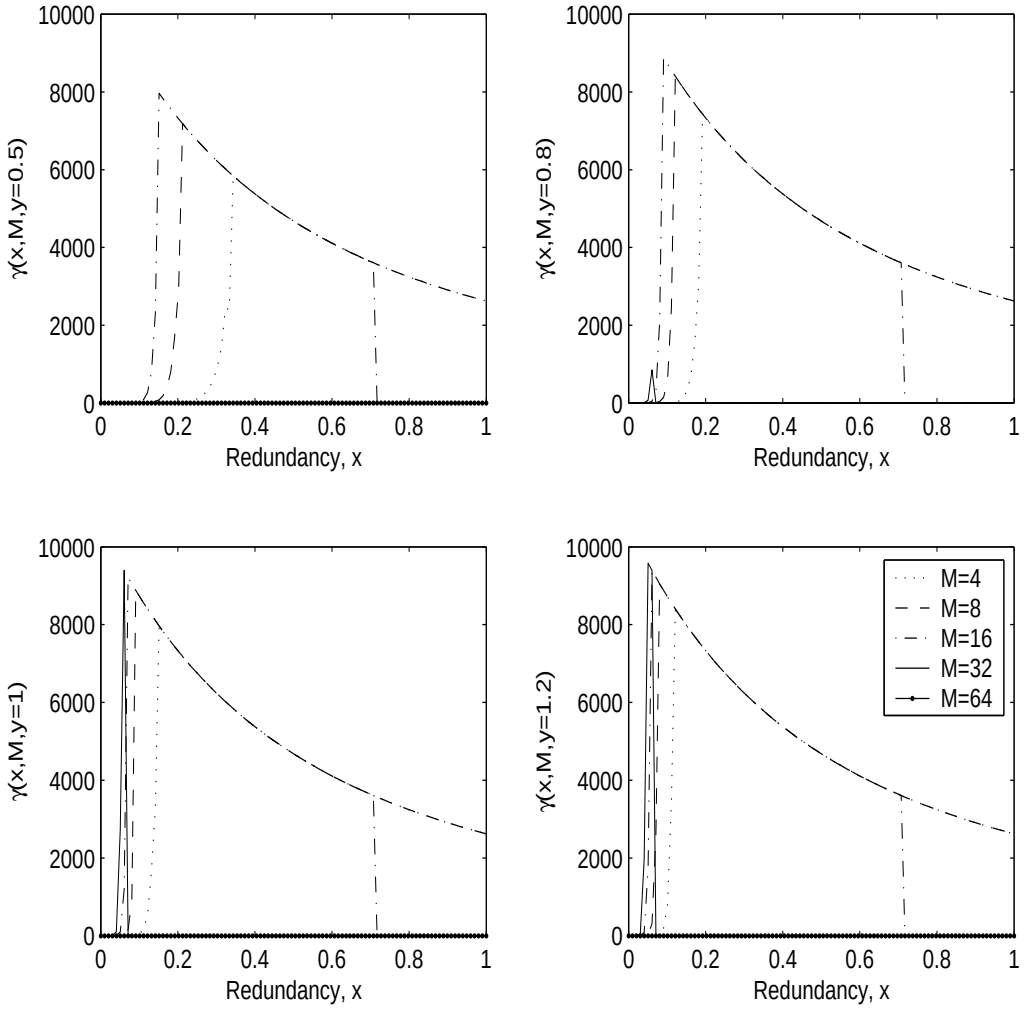
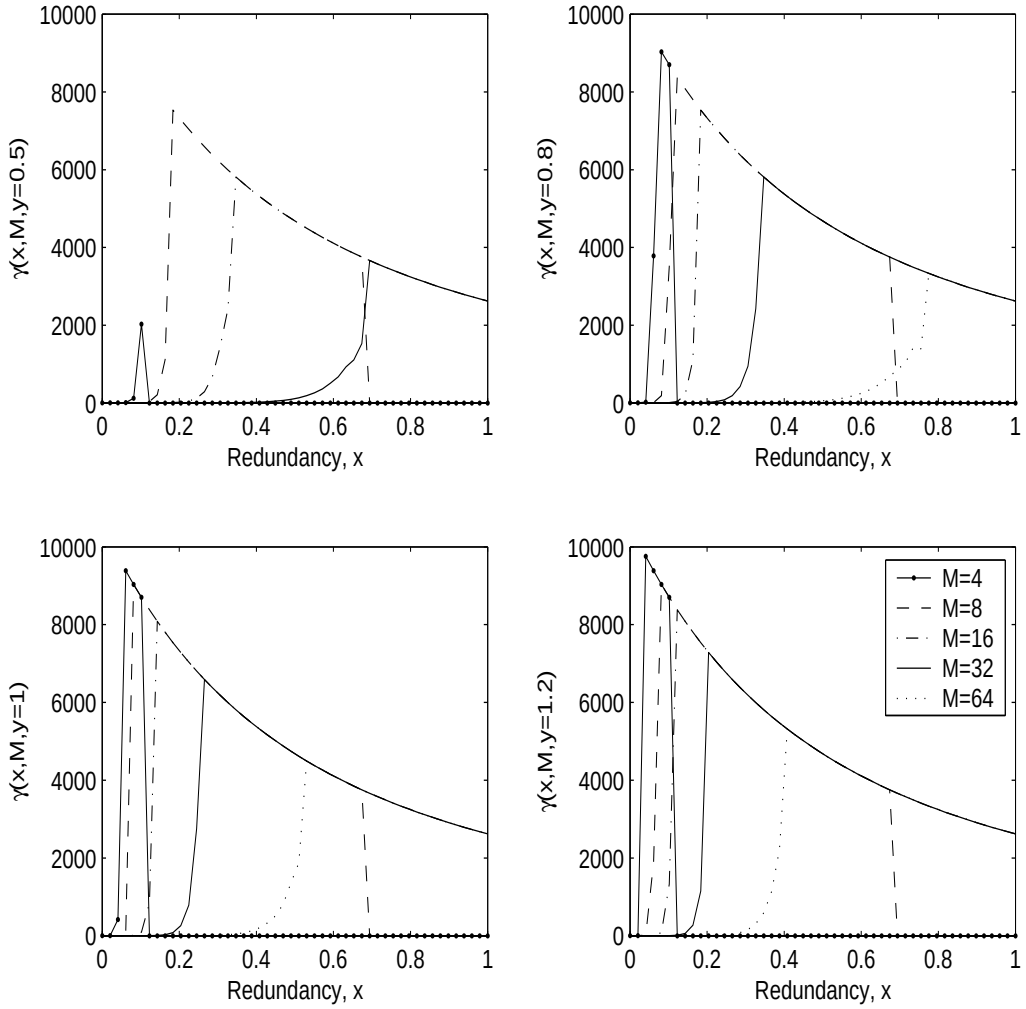


Figure 7.11: MQAM with bandwidth restrictions



Summarizing, it is possible to increase the optimum value of the objective function by:

1. Increasing the number of symbols, M, if the modulation format is MFSK.
2. Decreasing the number of symbols, M, if the modulation format is MQAM or MPSK.
3. Increasing the transmission power.
4. Increasing the redundancy.

Of course each one of these solutions has its drawbacks:

- Solutions 1. and 2. are related to bandwidth occupation in the sense that is not possible to stretch these solutions without meeting bandwidth constraints.
- For what concerns the solution 3., even if it is not considered in the simulations, it is not possible to increase the transmission power over a certain value in wireless links because of interference problems.
- Solution 4. is useful as long as the FEC has not completely cleaned the link from TCP point of view, 7.1.

Chapter 8

Ideas for future works

8.1 Case of fading channel

As it has been shown in section 2.1, wireless systems have to deal with a transmission channel that has several unwanted properties:

- Additive white gaussian noise.
- Frequency selectivity due to multipath propagation.
- Time variance of the channel due to Doppler spread also called fast fading.
- Multiple access interference (MAI) in CDMA-systems.

It has to be noted that in the previous simulations only one user was considered and the channel was supposed to be affected only by white gaussian noise. This model of the channel is appropriated for a wide range of physical channels but it results to be inadequate in characterizing signal transmission over radio channels whose transmission characteristics change with time.
It would be interesting to examine the problem of objective function maximization considering multiple users and fading interferences caused by multipath.

It has to be noted that considering a nonselective fading channel [1], performances of binary modulations are represented in figure 8.1.

[1] for more information see appendix A

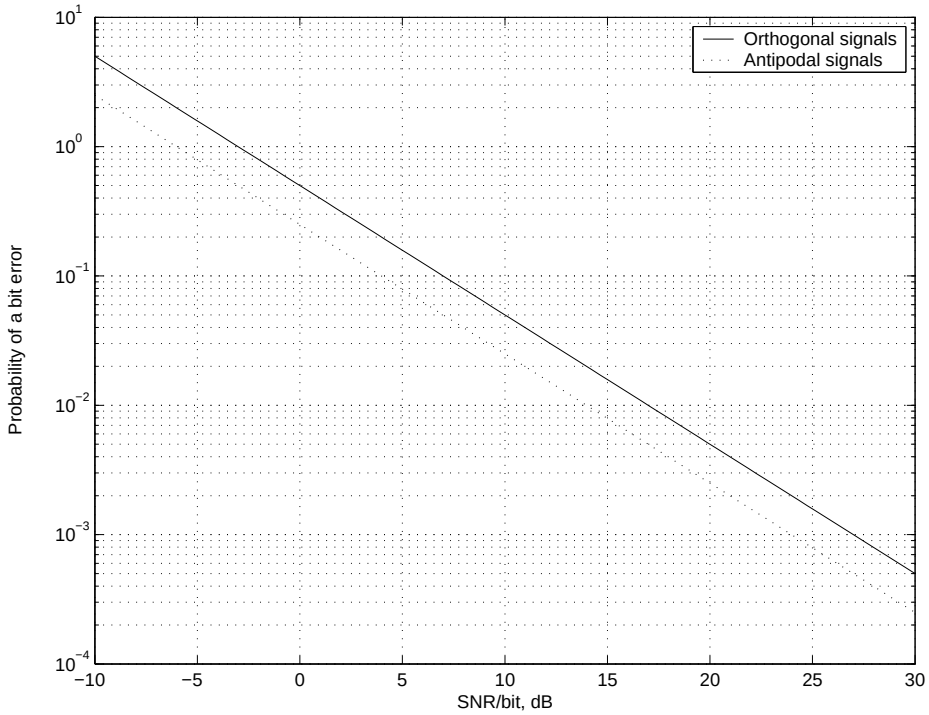


Figure 8.1: performance of binary signaling on a Reyleigh fading channel.

The trends shown in that figure result to be in contrast with the exponential trends of the same modulations format in the AWGN channel. This implies that in a fading channel , the transmitter must transmit with a higher power in order to obtain the same error probability. Also in this case the solution would be the use of redundancy that is performed through the use of diversity techniques.

If binary PSK is the transmission method and one of the possible diversity techniques is used, it can be demonstrated that an approximation of the error probability is the following, [4]:

$$p_e \approx \left(\frac{1}{4\overline{\gamma}_c}\right)^L \begin{pmatrix} L-1 \\ L \end{pmatrix} \tag{8.1}$$

where L is the number of diversity channels and $\overline{\gamma}_c$ is the average SNR per channel and it can be expressed as in [17]:

$$\overline{\gamma}_c = \frac{1}{1-\mu} \tag{8.2}$$

where μ is derived as a function of the signal to interference ratio as in [17]. The expression of the error probability is obtained assuming the use of an optimum combiner called *maximal ratio combiner*. The realization of this optimum combiner is based on the assumption that the channel attenuation and the phase shifts are known perfectly, that means that they are estimated without noise. From equation 2.15 it is clear that the probability of error varies as $1/\gamma_c$ raised to the Lth power and so the error rate decreases inversely with the Lth power.
Similar expressions can be found for coherent and noncoherent FSK and for differential PSK.

8.1.1 Proposals for TCP throughput evaluation over a fading multiple channel

The idea is to consider the formulas used for the throughput in the previous chapters by substituting the error probability found for the fading channel (i.e., expression 8.1 could be used in formula 6.7). Considering, for instance, the error probability for BPSK with diversity, the value of $\overline{\gamma}_c$ has to be specified. In that case it would be interesting to take into account the joint rate and power adaptation technique proposed in [17]. In fact the environment

considered in [17] is much more general than the one used for this report because it takes into account the interference problems caused by multiple users and it models the radio propagation by path loss, shadowing and multipath fading (this last problem is moreover generalized by considering also the case of multipath fading with unequal path power).
The problems presented by both the two assumptions are faced by the use of joint rate and power allocation, for both uniform and nonuniform traffic, basing on two different algorithms called DFA and BFA. This procedure allows the definition of $\overline{\gamma}_c$ to be substituted in the expression of the error probability.
In fact, assuming that the number of MSs in cell j is g_j, it is shown that the achieved downlink SIR, $(\mathrm{SIR})_{o,d}^{(j)}$ corresponding to all MSs in cell j can be written as [21]:

$$(\mathrm{SIR})_{o,d}^{(j)} = \frac{N}{\sum_{i=1}^{g_j} m_i^{(j)} \mu_i^{(j)}} \left(1 - \frac{P_c}{P_{B,j}}\right) \tag{8.3}$$

where:

- $m_i^{(j)}$ are the normalized transmission rates of all MSs in cell j.
- $P_{B,j}$ is the total transmission power of the jth BS.
- P_c is the pilot signal transmission power.

Both $m_i^{(j)}$ and $P_{B,j}$ can be determined basing on rate adaptation and power allocation.
It would be possible to examine the throughput achieved by TCP in the case of multiple fading channel, joint rate and power adaptation, hybrid ARQ at the link layer, different modulation formats.

Chapter 9

Conclusions

In this thesis performance of TCP over wireless links have been studied for different modulation formats and different error recovery techniques. Basing on [7], an analytical framework has been developed and an *objective function* has been defined. This function is assumed to be dependent on the amount of redundancy added to the blocks to be transmitted through the wireless link, on the number of retransmission ARQ at the link layer can perform before leaving to TCP the correction of the block on an end-to-end basis (called persistency) and on the transmission power. The behavior of the objective function has been studied for varying redundancy, transmission power and persistency. Moreover the analytical framework has been applied to relevant case studies basing on different modulation formats: GMSK, DBPSK, GFSK, MPSK, MQAM and MFSK. The constraint due to a limited amount of available bandwidth has been finally taken into account.
The results obtained show that using an hybrid-ARQ of the first type instead of a FEC alone to recover errors, increases the maximum value achieved by the objective function by using a lower value of redundancy. From the first simulations it is also established that an increasing in the transmission power improve link reliability by allowing a higher value of the objective function for a lower value of the redundancy.
Afterwards, different simulations have been done for different modulation formats and different results have been obtained. Considering at first the absence of bandwidth constraints, the MFSK format results to be the more convenient modulation scheme because it is possible to achieve higher values of the objective function by increasing the number of symbols and maintaining a constant value of the redundancy; while, for MQAM and MPSK

formats, an increase in the number of symbols does not cause any improvement.
On the other hand, while considering bandwidth constraints, it has been shown that for MFSK it is not possible to use too high values of the number of symbols because such a modulation format requires a higher bandwidth than the available one. Simulations for MPSK and MQAM with bandwidth constraints have shown that it is not always possible to use a high value of the number of symbols M (the ideal value would be $M = 2$) because the bandwidth required by this modulation format would overcome the available one by reducing the maximum value of the objective function and so the bind solution to achieve the maximum possible value of the objective function in this case, is to switch to a higher value of the number of symbols.
To conclude, a lot of possible solutions can be found to improve the objective function but they are strongly dependent on the particular environment and especially on the available bandwidth, the transmission power and the cost of different modulation formats.

For what concerns ideas for future works, basing on the previous observations, it would be interesting to implement a software that calculates what are the optimum modulation schemes and values of redundancy to be used in a certain environment. Moreover an ideal modulation and coding adaptation would be implemented by the analysis of the varying parameters of a general wireless channel (fading channel), [17]. Finally, it would be interesting to implement an hybrid-ARQ II type instead of the I type used in this work and to assume the transmission of the blocks on a back-to-back way for the computation of the RTT.

Appendix A

Channel model in case of fading

A radio channel has two main characteristics: the first one is due to the different time delays associated to each multiple propagation path, while the second is related to the variations that may appear in the structure of the medium. For the first reason the channel is said to be *time dispersive*, [16]. The *Doppler frequency spread* is a measure of how rapidly a received signal may vary with time.
The amplitude variation in the received signal due to multipath propagation are usually called *signal fading*.
A general model for a time variant multipath channel is shown in figure A.1. It consists in a tapped delay line with uniformly spaced taps that are modelled as complex-valued, Gaussian random processes that are mutually uncorrelated. The taps are spaced by $1/W$, where W is the bandwidth of the signal transmitted through the channel. If the amplitude of these complex variables can be described by a Reyleigh probability distribution, than the channel is called a *Reyleigh fading channel*.

Besides multipath time spread, T_m and Doppler frequency spread, D_m there are two additional parameters that are useful to characterize fading multipath channels:

- *Coherence time*: the reciprocal of the Doppler spread

$$T_{ct} = \frac{1}{2B_d} \tag{A.1}$$

- *Coherence bandwidth*: the reciprocal of the muultipath spread; it is a measure of the level of correlation of signal components, in other words,

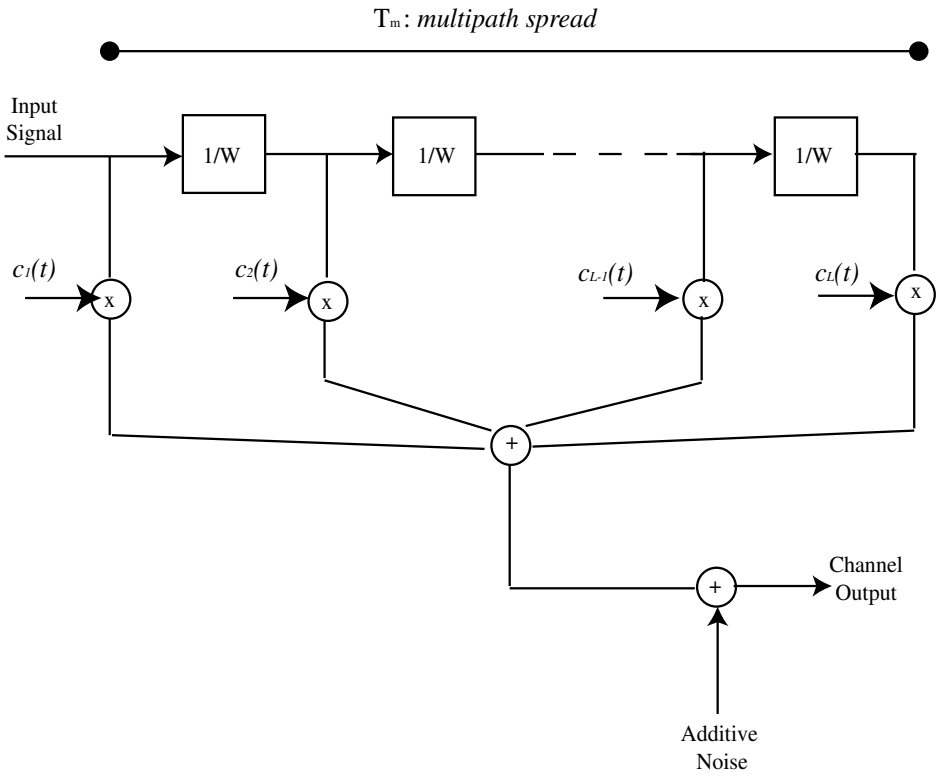


Figure A.1: Model for time variant multipath channel.

all signals within this bandwidth are affected by the same distorsions.

$$B_{cb} = \frac{1}{2T_m} \tag{A.2}$$

When the transmitted signal bandwidth W is larger than the coherence bandwidth B_{cb}, the components of the signal are attenuated and phase shifted differently by the channel and the channel is said to be *frequency selective*. On the other hand if the bandwidth of the signal is small enough comparing to the coherence bandwidth, the channel is said to be *frequency-nonselective*.

A.1 Performance of binary modulation over a nonselective slowly fading channel

In the case of nonselective channel, its impulse response can be represented as follows:

$$h(\tau; t) = \alpha(t)\delta(\tau - \tau_0(t)) \tag{A.3}$$

where α(t) has a Reyleigh distribution at any instant in time and the channel is said to be slow because the time variations of α(t) and τ_0 are very slow compared to the symbol interval.
It is possible to demonstrate that the probabilities of error for binary antipodal and orthogonal signals are the following:

$$P_e \approx \frac{1}{4\overline{\rho_b}} \qquad \text{antipodal signals} \tag{A.4}$$

$$P_e \approx \frac{1}{2\overline{\rho_b}} \qquad \text{orthogonal signals} \tag{A.5}$$

$$P_e = \frac{1}{2(1+\overline{\rho_b})} \qquad \text{DPSK} \tag{A.6}$$

$$P_e = \frac{1}{2+\overline{\rho_b}} \qquad \text{noncoherent FSK} \tag{A.7}$$

where, by definition,

$$\overline{\rho_b} = \frac{E_b}{N_0}\mathrm{E}(\alpha^2) \tag{A.8}$$

and hence $\overline{\rho_b}$ is the average received SNR/bit and $\mathrm{E}(\alpha^2)=2\sigma_\alpha^2$, where σ_α^2 is a parameter that characterize the Reyleigh distribution. The trend of the first two error probabilities presented are shown in figure 8.1.

A.2 Performance of binary modulation over a selective slowly fading channel

See section

A.3 Diversity Techniques

The idea of the diversity technique is based on the concept that errors occur in the channel when it is suffering a large attenuation. In this case if the receiver gets, over independently fading channels, several replicas of the signals transmitted, it would be able to reconstruct the original signal.

- *Frequency Diversity* With this method the signal is transmitted several times using several carriers, which are separated in frequency of an amount that is equal or bigger than the coherence bandwidth.
- *Time Diversity* Another solution to transmit the signal several times is to put its replicas in several time slots, where the separation between each time slot has to be equal or bigger than the coherence time of the channel.
- *Rake matched filter* It is the optimum receiver for processing a wideband signal used to provide the receiver with several independently fading signal paths.

- Finally there are other diversity techniques such like: use of multiple antennas, angle of arrival diversity, polarization diversity.

Appendix B

Cellular communication basics

A common feature of the previous mentioned mobile radio systems is that communication takes place between a stationary BS (Base Station) and a number of roaming MT (Mobile Terminal). Both the station and the terminal are expected to provide a sufficiently high signal level for the far-end receivers in order to maintain the required communication integrity. This is usually ensured by power control. The geographical area in which this condition is satisfied is called *traffic cell* and it typically has an irregular shape. In an ideal situation the total bandwidth amount for a specific mobile communication can be allocated within each cell, assuming that there is no energy split in the adjacent cells. But as the wave propagation cannot be shielded at the cell boundary, MSs near the cell edge would experience approximately the same signal energy within their channel bandwidth from at least two BSs. This phenomenon is called *cochannel interference*. A solution to this problem is to divide the total bandwidth $B_{cell} = B_{tot}/N$ in frequency slots and assigning each part of this bandwidth to each traffic cell within a *cluster* of N cells. The drawback of this solution is that the total number of MSs that can be supported by each cell is reduced by a factor of N. One of the possible solutions to this problem would be to make the cluster as small as the original cell and this is achieved by reducing the transmission power. Another advantage presented in this solution is the improvement in the propagation environment due to both the presence of a dominant line of sigh propagation path and the mitigated effects of the multipath propagation (for further information see chapter 4).
Another important issue associated with the cellular concept and cellular planning is DCA, the dynamic channel assignment algorithm. In a dynamic

assignment method, all channels are potentially available to all cells and are assigned to cells dynamically as calls arrive. If this is done right, it can take advantage of temporary changes in the spatial and temporal distribution of calls in order to serve more users. The family of DCA algorithm is closely related to a range of multiple access schemes that are considered in the following section.

B.1 Multiple Access

If the physical channels assigned to a MT through a given frequency slot for the entire call, this means that *Frequency Division Multiple Access* is the technique used. FDMA was very common in most first generation mobile radio systems.
In second generation systems such as the pan-European GSM [22], the American IS-54 [24], *Time Division Multiple Access* was purposed by assigning the whole bandwidth of a TDMA carrier to one MS for a given interval of time (the duration of a time slot).
The American IS-95, [23], *Code Division Multiple Access* (CDMA) system uses all the system bandwidth all the time for all users, using a *Spread Spectrum Technique*. Individual conversations are encoded with a pseudo-random digital sequence.

Bibliography

[1] D. Barman, I. Matta, E. Altman, R. E. Azouzi, "TCP Optimization through FEC, ARQ and Transmission Power Tradeoffs", Technical Report, December 3, 2003.

[2] J. Padhye, V. Firoiu, D. Towsley, J. Kurose, "Modeling TCP Throughput: A Simple Model and its Empirical Validation",in *Proceeding of ACM/SIGCOMM 98*, Vancouver, Canada, October 1998.

[3] L. Galluccio, G. Morabito, S. Palazzo,"An Analytical Study of a Tradeoff between Transmission Power and FEC for TCP optimization in Wireless Networks".

[4] J. G. Proakis,"Communication Systems Engineering", *Prentice Hall International Editions*,1994.

[5] Bernard Sklar,"Defining, Designing, and Evaluating Digital Communication Sistems", IEEE Communications magazine, November 1993.

[6] M.Luby, L. Vicisiano, J. Gemmell, L. Rizzo, M. Handley, J. Crowcroft, "RFC 3453 (rfc3453)-The Use of Forward Error Correction", December 2002.

[7] S. Lin, D. J. Costello jr., "Error Control Coding: Fundamental and Applications", *Prentice Hall International Editions*,1983.

[8] F. Graziosi, M. Pratesi, "Dispense del corso di sistemi di radiocomunicazione", A.A. 2003/2004.

[9] Jean-Paul M. G. Linmartz, "Wireless Communication, The Interactive Multimedia CD-ROM", Baltzer Science Publishers, P.O.Box 37208, 1030 AE Amsterdam, ISSN 1383 4231, Vol. 1 (1996), No.1.

[10] C. Fischione,"TCP and Radio Resource Management", Draft, July 14 2004.

[11] J. Postel, "RFC 792 (RFC792) - Internet Control Message Protocol", September 1981.

[12] Information Sciences Institute University of Southern California," RFC 792 (RFC792)-Transmission control Protocol",Internet RFC/STD/FYI/BCP Archives, September 1981.

[13] Egli, "Radio Propagation Above 40MC Over Irregular Terrain", (Proceedings of the IRE, Vol. 45, Oct. 1957, pp.1383-1391).

[14] Sklar, B., Digital Communications: Fundamentals and Applications, Second Edition (Upper Saddle River, NJ: Prentice-Hall, 2001).

[15] 3GPP TS 25.214 V6.1.0, "3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Physical layer procedures (FDD)".

[16] J. G. Proakis, M. Salehi, "Communication Systems Engineering", Prentice Hall International Editions, 1994.

[17] E. Hossain, D. I. Kim, B. K. Bhargava, "Analysis of TCP Performance Under Joint Rate and Power Adaptation in Cellular WCDMA Networks", ieee transactions over wireless communications, vol. 3, no. 3, May 2004.

[18] W. Stevens, "TCP/IP illustrated", vol.1 "The Protocols", Addison-Wesley, 1994.

[19] W. Stevens, "TCP Slow Start, Congestion Avoidance, Fast Retransmit, and Fast Recovery Algorithms", RFC2001, January 1997.

[20] C. Barakat, E. Altman, "Bandwidth tradeoff between TCP and link-level FEC", Computer networks 39 (2002) 133-150, November 2001.

[21] D. I. Kim, E. Hossain, V. K. Bhargava, "Downlink joint rate and power allocation in cellular multi-rate WCDMA systems", IEEE trans. wireless commun., vol.2, pp.69-80, January 2003.

[22] GSM Recommendation 05.05, Annex 3, pp. 13-16, November 1998.

[23] Public Digital Cellular (PDC) Standard, RCR STD-27.

[24] Dual-mode subscriber equipement- Network Equipement Compatibility Specification, Interim Standard IS-54, Telecommunications Industry Assosiation, 1998.

[25] C. Berrou, A. Glavieux, P. Thitimasjshima, "Near Shannon Limit Error-Correcting Coding and Decoding: Turbo Codes", proceeding of the IEEE International Conference on Communications (May 1993, Geneva, Switzerland): 1064-70.

[26] J. Postel,J. Reynolds, "RFC 854 - Telnet Protocol Specification", ISI, 18639 May 1983.

[27] J. Postel,J. Reynolds, "RFC 959 - File Transfer Protocol", ISI, october 1985.

[28] C. Fischione, F. Graziosi, F. Santucci, N. Mller, K. H. Johansson, H. Hjalmarsson, "Modelling of TCP over wireless under Power Control, MAI, Link Level Error Recovery and Mixed Traffic Sources".

[29] N. Mller, K. H. Johansson, H. Hjalmarsson, "Making retransmission delays in wireless links friendlier to TCP", 43rd IEEE Conference on Decision and Control December 14-17, 2004 Atlantis, Paradise Island, Bahamas.