

Wireless Communications Systems and Networks

Edited by
Mohsen Guizani

Wireless Communications Systems and Networks

Information Technology: Transmission, Processing, and Storage

Series Editor: **Jack Keil Wolf**

*University of California at San Diego
La Jolla, California*

Editorial Board: **Robert J. McEliece**

*California Institute of Technology
Pasadena, California*

John Proakis

*Northeastern University
Boston, Massachusetts*

William H. Tranter

*Virginia Polytechnic Institute and State University
Blacksburg, Virginia*

Coded Modulation Systems

John B. Anderson and Arne Svensson

Communication System Design Using DSP Algorithms: With Laboratory

Experiments for the TMS320C6701 and TMS320C6711

Steven A. Tretter

A First Course in Information Theory

Raymond W. Yeung

Interference Avoidance Methods for Wireless Systems

Dimitrie C. Popescu and Christopher Rose

Nonuniform Sampling: Theory and Practice

Edited by Farokh Marvasti

Simulation of Communication Systems, Second Edition: Methodology, Modeling, and Techniques

Michel C. Jeruchim, Philip Balaban, and K. Sam Shanmugan

Stochastic Image Processing

Chee Sun Won and Robert M. Gray

Wireless Communications Systems and Networks

Mohsen Guizani

A Continuation Order Plan is available for this series. A continuation order will bring delivery of each new volume immediately upon publication. Volumes are billed only upon actual shipment. For further information please contact the publisher.

Wireless Communications Systems and Networks

Edited by

Mohsen Guizani

*Western Michigan University
Kalamazoo, Michigan*

KLUWER ACADEMIC PUBLISHERS
NEW YORK, BOSTON, DORDRECHT, LONDON, MOSCOW

eBook ISBN: 0-306-48642-3
Print ISBN: 0-306-48190-1

©2004 Kluwer Academic Publishers
New York, Boston, Dordrecht, London, Moscow

Print ©2004 Kluwer Academic/Plenum Publishers
New York

All rights reserved

No part of this eBook may be reproduced or transmitted in any form or by any means, electronic, mechanical, recording, or otherwise, without written consent from the Publisher

Created in the United States of America

Visit Kluwer Online at: <http://kluweronline.com>
and Kluwer's eBookstore at: <http://ebooks.kluweronline.com>

This work is dedicated to
my parents, my wife Saida,
and my kids Nadra, Fatma, Maher, Zainab, Sara, and Safa

This page intentionally left blank

ACKNOWLEDGMENT

This work would have not been possible without the help and encouragement of so many people. These people inspired my knowledge and helped directly or indirectly with many aspects of the writing process, as well as ideas of topics of interest to include in this book. These include, Ibrahim Habib, City University of New York, Ted Rappaport, University of Texas-Austin, K. C. Chen, National Taiwan University, and H-H. Chen, National Sun-Yat Sen University in Taiwan, to name a few. I am also very thankful to my graduate students Ghassen Ben Brahim, Mohamed Bourouha, and Anupama Raju for their help as well as Kluwer editors Ana Bozicevic and Alex Greene who facilitated many processing issues to get this work finished at a fast pace.

I also would like to thank all the contributors to the chapters—their patience during this process was very much appreciated.

This work is dedicated to my parents, my wife Saida and my kids Nadra, Fatma, Maher, Zainab, Sara, and Safa for their patience and support during the time I spent working on this project.

This page intentionally left blank

PREFACE

Since the early 1990s, the wireless communications field has witnessed explosive growth. The wide range of applications and existing new technologies nowadays stimulated this enormous growth and encouraged wireless applications. The new wireless networks will support heterogeneous traffic, consisting of voice, video, and data (multimedia). This necessitated looking at new wireless generation technologies and enhance its capabilities. This includes new standards, new levels of Quality of Service (QoS), new sets of protocols and architectures, noise reduction, power control, performance enhancement, link and mobility management, nomadic and wireless networks security, and ad-hoc architectures. Many of these topics are covered in this textbook.

The aim of this book is research and development in the area of broadband wireless communications and sensor networks. It is intended for researchers that need to learn more and do research on these topics. But, it is assumed that the reader has some background about wireless communications and networking. In addition to background in each of the chapters, an in-depth analysis is presented to help our readers gain more R&D insights in any of these areas. The book is comprised of 22 chapters, written by a group of well-known experts in their respective fields. Many of them have great industrial experience mixed with proper academic background.

The first chapter discusses peer-to-peer networking in mobile communications based on SIP. J. Huber of Siemens, Germany, describes the evolution path of the Global System for Mobile Communications towards UMTS that integrates Internet technologies and provides a platform for the development of new mobile specific services. These services include new developments that will incorporate mobility and person-to-person communications with real-time video, sound, and voice, describing the evolutionary path in achieving this objective.

The second chapter, written by A. Boukerche from the University of Ottawa, Canada, discusses some protocols for data propagation in wireless sensor networks. These protocols cover large numbers of these sensors that can be deployed in areas of interest, such as inaccessible terrains or disaster places, and

use self-organization and collaborative methods to form a network. Their wide range of applications is based on the possible use of various sensor types. Thus, sensor networks can be used for continuous sensing, event detection, location sending, as well as micro-sensing. This is a challenging technological and algorithmic task. Features, including the huge number of sensor devices involved, the severe power shortage, computational and memory limitations, their dense deployment and frequent failures, pose new design analysis and implementation challenges. This chapter aims to present certain important aspects of the design, deployment, and operation of sensor networks, as well as some evaluation of their performance analysis.

The third chapter, by A. Safwat of Queen's University, Canada, talks about the design of a novel architecture for 4G and 4G+ wireless networks called "A-CELL." The A-Cell relay is used to reduce power consumption and enhance coverage and throughput. A-Cell increases spatial reuse via directive antennas and GPS. The author derives an analytical model for A-Cell based on multi-dimensional Markov chains. The model is then used to formulate A-Cell call blocking. This is the first time that directive antennas are used as a means to reduce interference, conserve energy, and enhance spatial reuse in a multi-hop UMTS wireless network.

H. Holma and A. Toskala from Nokia, Finland discuss third generation WCDMA radio evolution. They discuss technologies including: 2 Mbps data rate with variable bit rate capability, support of multi-service, QoS differentiation, and efficient packet data. These technologies also contain downlink packet data operation enhancement, high speed downlink packet access (HSDPA). HSDPA utilizes Hybrid ARQ and higher order modulation for improving data spectral efficiency and for pushing bit rates beyond 10 Mbps. The further 3GPP releases will study the enhancements of packet data performance in uplink. Other important features in future 3GPP releases include advanced antenna technologies and WCDMA standard for new spectrum allocations. The chapter also describes the main solutions of 3GPP WCDMA standard in more detail.

Next, H-H. Chen from National Sun Yat-Sen University, Taiwan, discusses the evolution of CDMA from interference-limited to noise-limited. The chapter covers various new CDMA-based wireless standards—such as TD-SCDMA and LAS-CDMA, etc., as well as those on other commonly referred CDMA-based standards, such as IS-95A/B, cdma2000, UTMS-UTRA, W-CDMA, etc.—providing a rather comprehensive coverage of the evolution of CDMA technologies from interference-limited to noise-limited. The discussions on various features of state-of-the-art CDMA technologies lead to the introduction of a new concept on *isotropic CDMA air-link technology*, which consists of two related sectors: isotropic spreading codes and isotropic spreading modulation. This can offer a homogenous performance in both synchronous and asynchronous CDMA channels. Based on the discussion on conventional CDMA technologies and introduction of two novel CDMA architectures, the author concludes that the

evolution of CDMA technologies will eventually make a CDMA system offer a noise-limited performance.

Then, B. Hashem from Nortel Networks of Ottawa, Canada, describes a power control implementation in third generation CDMA systems. The inner loop and the outer loop algorithms are explained whereby the power commands are sent at a high rate, and uncoded, resulting in high error rate in receiving them. The problem of errors in the power control commands becomes worse during soft handoff where the terminal communicates simultaneously with more than one base station. A few algorithms are introduced during soft handoff to increase the system capacity, like power control rate reduction and the adjustment loop. Adjusting the transmission rate is also discussed as another scheme to achieve the required quality.

F. Gunnarsson of Linköping University, Sweden, discusses the characteristics and fundamentals of power control in wireless networks. He describes power control fundamentals, including both theoretical and practical limitations. The relationship to session management, such as admission and congestion control, is also addressed. Concepts and algorithms are illustrated with simple examples and simulations.

L. Yang and M-S. Alouini of the University of Minnesota present their results on the average outage duration of wireless communication systems. More specifically they discuss generic results for the LCR and AOD: (1) of maximal ratio combining systems subject to CCI operation over independent identically distributed Rician and/or Nakagami fading environments when a minimum desired signal power requirement is specified for satisfactory reception; and (2) of various selection combining diversity scheme in presence of multiple CCI and with both minimum SIR and desired signal power constraints over independent, correlated, and/or unbalanced channels. Corresponding numerical examples and plots illustrating the mathematical formulation are also provided and discussed.

A transport level approach in enhancing TCP performance in wide area cellular wireless networks are discussed by E. Hossain and N. Parvez from the University of Manitoba, Canada. They present a comprehensive study on the performances of the basic TCP variants (e.g., TCP Tahoe, TCP Reno, TCP New-Reno, SACK TCP, FACK TCP) in wide-area cellular wireless networks. For the basic TCP variants, an in-depth analysis of the transport-level system dynamics is presented based on computer simulations using *ns-2*. Impacts of variations in wireless channel error characteristics, number of concurrent TCP flows, and wireless link bandwidth on the average TCP throughput and fairness performances are investigated. The maximum achievable throughput under window-based end-to-end transmission control is also evaluated and the throughput performances of TCP New-Reno, SACK TCP, and FACK TCP are compared against this ideal TCP throughput performance. To this end an overview of the major modifications to the basic TCP variants based on transport-level approaches to enhance TCP performance in wireless networks is presented.

The deployment of several real-time multimedia applications over the Internet has motivated a considerable research effort on the provision of multiple services on the Internet. In order to extend this work over wireless links however, we must also take into account the performance limitations of wireless media. G. Xylomenos and G. Polyzos from the Athens University of Economics and Business in Greece survey various related approaches and conclude that link layer schemes provide a universal and localized solution. Based on simulations of application performance, over many link layer schemes, they show that different approaches work best for different applications by concurrently supporting multiple link layer schemes. Simulations of multiple applications executing simultaneously show that this approach dramatically improves performance for all of them. They finally consider embedding this approach into a Quality of Service oriented Internet, discussing the traditional best-effort architecture, the Differentiated Services architecture, and an advanced dynamic service discovery architecture.

A value added service to broadband wireless network is the remote access virtual private network (VPN). The corporate legitimate portable users can connect to their users through a wireless network from different locations and get secure services as if they were connected to the corporate local area network. One of the main challenges is to block illegitimate wireless users' requests. Registration and authentication functions should be implemented with highly secured wireless connections. These functions are accomplished by tunnelling the user information, in a secured form, to the corporate authentication server through the Internet traffic. The Corporate Authentication Server then grants or denies the user access. M. Matalgah from the University of Mississippi, J. Qaddour, from Illinois Institute of Technology, and their colleagues O. Elkeelany and K. Sheikh from Sprint Inc. have designed a portability architecture for nomadic wireless Internet access users and security performance evaluation to solve these challenges. This chapter addresses various portability scenarios, architectures, implementation, and requirement issues for portable wireless Internet access systems. Moreover, performance evaluation and comparison are presented for the state-of-the-art security and authentication techniques.

V. W-S Feng and Y-B Lin, of National Chiao Tung University and Industrial Technology Research Institute (ITRI) of Taiwan, describe the design and implementation of a softswitch for Third Generation UMTS Mobile All-IP Network developed by Computer and Communication Research Laboratory (CCL) of ITRI. This softswitch can be utilized as call agent (media gateway controller) for the third generation mobile all-IP network such as UMTS. The CCL Softswitch follows the reference architecture proposed by International Softswitch Consortium. In this approach, the Intelligent Network (IN) call model is implemented to interwork with existing IN devices. They design protocol adapter modules and service provider interface to ensure that multiple VoIP protocols can be supported without modifying the core of the softswitch.

Furthermore, the message flows of call setup, call transfer, and inter-softswitch call are described to show the feasibility of the softswitch.

Then, in Chapter 13, E. Hossain et al., from the University of Manitoba, Canada, discuss issues and approaches of clustering in mobile wireless ad hoc networks. In a multi-hop ad hoc wireless network, which changes its topology dynamically, efficient resource allocation, energy management, routing, and end-to-end throughput performance can be achieved through adaptive clustering of the mobile nodes. Impacts of clustering on radio source management and protocol performance in a multi-hop ad hoc network are described and a survey of the different clustering mechanisms is presented. A comparative performance analysis among the different clustering mechanisms based on the metrics such as cluster stability, load distribution, control signaling overhead, energy-awareness is performed.

The maximum capacity of a CDMA cellular system's radio interface depends on the time varying radio environment. This makes it hard to establish the amount of currently available capacity. The received interference power is the primarily resource in the uplink. Ability to predict how different resource management decisions affect this spatial quality is therefore of utmost importance. The uplink interference power is related to the uplink load through the pole equation. In this chapter, E. Lundin and F. Gunnarsson of Ericsson Research and Linköping University, Sweden, discuss both theoretical and practical aspects of uplink load estimation concentrating on concepts and algorithms.

In Chapter 15, end-to-end performance analysis of multi-hop transmissions over Rayleigh fading channels is presented. Several types of relays for both regenerative and non-regenerative systems are considered. In addition, optimum power allocation over these hops is investigated as it is considered a scarce resource in the context of relayed transmission. Numerical results show that regenerative systems outperform non-regenerative systems especially at low average signal-to-noise ratios, or when the number of hops is large. The authors, M. Hasna and M-S. Alouini of Qatar University and the University of Minnesota, also show that power optimization enhances the system performance, especially if the links are highly unbalanced in terms of their average fading power or if the number of hops is large. Interestingly, they also show that non-regenerative systems with optimum power allocation can outperform regenerative systems with no power optimization.

Mobility management plays a significant role in current and future wireless mobile networks in effectively delivering services to the mobile users on the move. Many schemes have been proposed and investigated extensively in the past. However, most performance analyses were carried out either under simplistic assumptions on some time variables or via simulations. Recently, Y. Fang and W. Ma of the University of Florida have developed a new analytical approach to investigate the modeling and performance analysis for mobility management schemes under fairly general assumptions. In this chapter, they present the

techniques they have developed for this approach and summarize major results obtained for a few mobility management schemes such as movement-based mobility management, pointer forwarding scheme (PFS), Two-level location management scheme, and two-location algorithm (TLA).

Pervasive computing environments present entirely new set of challenges because of the fact that data may be acquired and disseminated at various stages within the system. Therefore, novel caching mechanisms are needed that take into account demand-fetched and prefetched (or pulled), as well as broadcast (or pushed) data. In addition, cached maintenance algorithms should consider such features as heterogeneity, mobility, interoperability, proactivity, and transparency that are unique to pervasive environments. Pervasive computing applications require continual and autonomous availability of ‘what I want’ type of information acquisition and dissemination in a proactive yet unobtrusive way. Mobility and heterogeneity of pervasive environments make this problem even more challenging. Effective use of middleware techniques—such as caching—can overcome the dynamic nature of communication media and the limitations of resource-poor devices. These and other challenging issues in this area are discussed by M. Kumar and S. Das, of the University of Texas at Arlington, in Chapter 17. Das *et al.* also present issues of security in wireless mobile and sensor networks in Chapter 18. In particular, they discuss security issues and challenges in wireless mobile ad hoc and sensor networks as well as in pervasive computing infrastructures. They also describe security protocols for such environments.

H-H. Chen discusses various issues on pulse shaping waveform design for bandwidth efficient digital modulations in Chapter 19. He starts with an overview on the evolution of pulse shaping technologies from early digital modulation schemes to recently emerging new carrier modulations such as quadrature-overlapped modulations. The impact of the pulse shaping technologies on both bandwidth-efficiency and power-efficiency, which are two essential merit parameters of a digital modem, are also discussed with an emphasis on shaping pulse design methodology as well as performance analysis of a pulse-shaped digital modem. Several new pulse shaped QO modulation schemes, which adopt various novel pulse shaping waveforms generated by time-domain convolution method, are presented and their performance analysis will be carried out in comparison with other traditional digital modems, such as QPSK, OQPSK and MSK.

In chapter 20, E. Biglieri and G. Taricco of Politecnico di Torino, Italy, investigate transmission systems where more than one antenna is used at both ends of the radio link. The use of multiple transmit and receive antennas allows one to reach capacities that cannot be obtained with any other technique using present-day technology. After computing these capacities, the authors show how “space-time” codes can be designed and how sub-optimum architectures can be employed to simplify the receiver.

In Chapter 21, a group of researchers (R. Schober, L. Lampe, S. Pasupathy, and W. Gerstacker) from the University of British Columbia and the University of Toronto, Canada, as well as the University of Wrlangen-Nuernberg, Germany, designed and optimized matrix-symbol-based space-time block codes (STBC's) for transmission over fading intersymbol interference (ISI) channels. They show that STBC's employing diagonal code matrices exclusively facilitate the successful application of suboptimum equalization techniques for the practically important case when only a single receive antenna is available. Three different types of diagonal STBC's are optimized for fading ISI channels and their performances are compared for decision-feedback equalization (DFE) and decision-feedback sequence estimation (DFSE), respectively. The robustness of the designed codes against variations of the characteristics of the fading ISI channel and the dependence of the equalizer performance on the STBC structure is investigated.

Last, but not least, M. Bassiouni, W. Cui, and B. Zhou of the University of Central Florida, present a fast routing and recovery protocols in hybrid ad-hoc cellular networks. They present and evaluate efficient recovery and routing protocols for mobile routers in HACN (hybrid ad-hoc cellular network). In the dual backup recovery protocol, a standby router is dedicated to each primary router. A more flexible protocol, called the distributed recovery protocol, is obtained by relaxing this one to one mapping and scattering the backup routers geographically among the primary routers. The effectiveness of the distributed recovery protocol is demonstrated by simulation results. A simple analytical model is also presented for computing the blocking probability of new calls in the presence of router failures. Chapter 22 is concluded by presenting an efficient location-based routing protocol for HACN mobile routers.

Mohsen Guizani

the multipath signal receiver, $(3, -1, 2, -2, 2, 1)$, is the received signal of transmitted bit stream after going through multipath channel and local flock correlators.

The proposed recursive multipath signal reception filter possesses several advantages. Firstly, it offers a very agile structure, the core of which is made up of two transversal filters; one for channel impulse response estimation and the other for data detection. Secondly, working jointly with the pilot signaling it performs very well in terms of the accuracy in channel impulse response estimation, as shown in Figure (5.19) and the obtained BER results followed. The multipath channel equalization and signal coherent-combining are actually implemented at the same time in the proposed scheme under a relatively simple hardware structure. Thirdly, it can operate adaptively to the channel characteristics variation without needing the prior knowledge of the channel, such as inter-path delay and relative strength of different paths, etc. On the contrary, a RAKE receiver in a conventional CDMA system requires every individual path gain coefficient for maximal ratio combining, which themselves are usually unknown and thus have to be estimated by resorting to other complex algorithms.

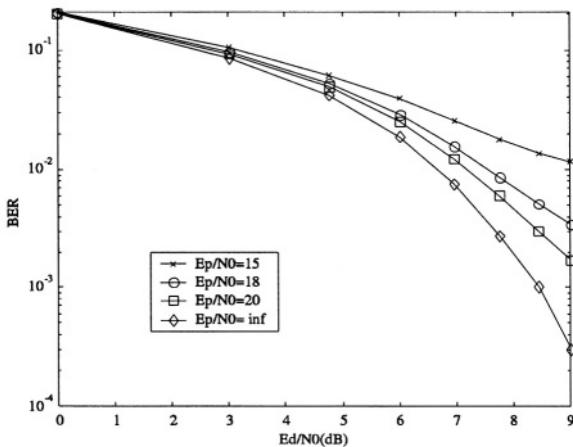


Figure 5.21. BER for CC-CDMA in down-link multipath channel using recursive filter, where pilot is used to estimate multipath parameters. Delay profile=[0.5774, 0.5774, 0.5774], inter path delay=1 chip, PG of CC code 4×16 , user number=4.

The performance of the CC-CDMA architecture with the recursive filter for multipath signal reception is shown in Figures (5.21) to (5.24), where two typical scenarios are considered: one for down-link performance and the other for up-link performance, similar to the performance comparison made for the MAI-AWGN channel concerned in Fig-

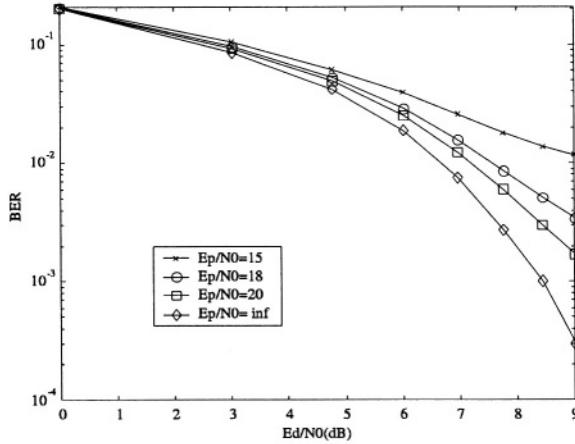


Figure 5.22. BER for CC-CDMA in up-link multipath channel using recursive filter, where pilot is used to estimate multipath parameters. Delay profile=[0.5774, 0.5774, 0.5774], inter path delay=1 chip, inter user delay=3 chip, PG of CC code 4×16 , and user number=4.

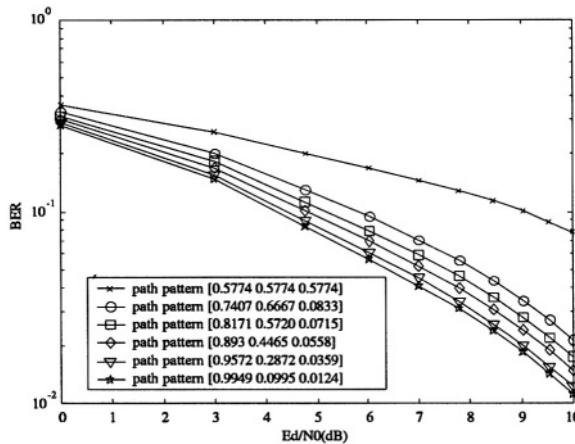


Figure 5.23. BER for CC-CDMA in down-link multipath channel using recursive filter, where pilot is used to estimate multipath parameters. Normalized multipath patterns are considered with inter path delay=1 chip, PG of CC code = 4×6 , pilot power=signal power, and user number=4.

ures (5.13) and (5.14). It is observed from the figures that the almost identical BER performance for both up-link and down-link channels is presented. The transmitting power of the pilot channel has a very strong influence on the BER of the CC-CDMA as a whole, as shown in Figures (5.21) and (5.22). Thus, in order to ensure a satisfactory performance,

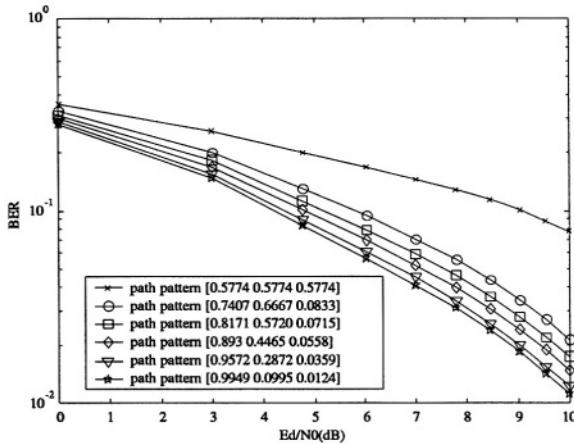


Figure 5.24. BER for CC-CDMA in up-link multipath channel using recursive filter, where pilot is used to estimate multipath parameters, Normalized multipath patterns are considered with inter path delay=1 chip, inter user delay=3 chip, PG of CC code = 4×6 , pilot power=signal power, and user number=4.

E_p/N_o should be made at least 18 dB; otherwise the errors in channel estimation will seriously affect the data detection process followed. Figures (5.23) and (5.24) show the performance of a CC-CDMA system under various multipath channels, which consist of three multipath returns and whose delay profile are normalized to reflect the fact that the total transmitting power for different channels is fixed.

5.3.3 Pros and Cons for CC-CDMA.

The major characteristic features of the CC-CDMA system are summarized as follows.

- It can offer attractive isotropic MAI-free operation for wireless applications, which is hardly achievable by using any other CDMA technologies available at the moment.
- The use of the ISM (or offset stacking spreading modulation) can greatly increase the spreading efficiency (SE), which is defined as bit(s) per chip. It is noted that currently available 2-3G CDMA based systems [1]-[6] can offer an SE at merely about $1/N$ bit per chip, where N is the spreading factor of the systems; whereas the CC-CDMA can push the SE figure up to one bit per chip, about N times higher than that obtainable in the 2-3G systems. As the system bandwidth is determined by the chip-width of spreading-modulated signal, a greater SE means higher bandwidth efficiency.

Thus, the above example of the MAI-free IST has a great potential to increase the overall bandwidth efficiency of a wireless system.

- Some other advantages of the CC-CDMA system include its agility to perform rate-matching algorithm for multi-media services, the flexibility in supporting asymmetrical traffic in up-link and down-link channels, etc. Those who are interested in more detail information on the CC-CDMA and the generation of CC codes may refer to [69].

The CC-CDMA architecture has also its own technical limitations, which can be listed below.

- Although the CC-CDMA system does offer an isotropic MAI-free operation, it should use a relatively complex receiver (a dedicated recursive filter) to detect signal in multipath channel with the help of the pilot signaling, as shown in [69]. In order words, the CC-CDMA architecture proposed earlier [69] does not have isotropic Mi-free property, without which the performance improvement under multipath channel is limited.
- Furthermore, the CC-CDMA needs some multi-level digital carrier modulation, such as QAM, to fulfill carrier transmission due to its multi-leveled baseband signal formulation after offset stacked spreading modulation, which further complicate transceiver hardware and motivates us to search for some other better solutions. In fact, the high bandwidth efficiency of the CC-CDMA relies on the high detection efficiency of the modulation scheme. Unfortunately, due to the need of multi-level digital modulation scheme, the CC-CDMA system will still face the same problem associated with the reduced detection efficiency as that in a multi-level modem under noise and interference.

Another concern with the CC-CDMA system is that a relatively small number of users can be supported with a family of the CC codes. Take $L=64$ CC code family as an example. It is seen from Table (5.3) that such a family has only 4 flocks of codes, each of which can be assigned to one channel (for either pilot or data). If more users should be supported, long CC codes have to be used. On the other hand, the maximum length of the CC codes is in fact limited by maximal number of different baseband signal levels manageable in a digital modem, as mentioned previously. One possible solution to this problem is to introduce frequency-divisions on top of the code-divisions in each frequency band to

create more transmission channels. Yet another possibility is to use long code scrambling together with the short CC code spreading to increase the number of users, which can be supported in the same frequency band. However, doing so might incur some other concerns about the system complexity and bandwidth efficiency, etc. Therefore, the most worthwhile work is to search for some other complementary codes, which retain all advantages of CC codes but with a larger code-set size.

Therefore, we have some strong reason to find some more preferable CDMA systems possessing both isotropic MAI-free and isotropic MI-free properties. One possible solution is the combination of CC codes and direct-sequence spreading modulation, or CC/DS-CDMA scheme, which provides very attractive properties, such as isotropic MAI-free and isotropic MI-free, near-far effect resistance, etc., and it works on a very simple receiver hardware, a correlator or matched-filter, to facilitate user portable terminal miniaturization. The detail about the CC/DS-CDMA scheme will be covered in the sub-section followed.

As the last remark before ending this sub-section, it should be emphasized that ordinary orthogonal codes can never guarantee isotropic MAI-free operation. For instance, the application of Walsh-Hadamard sequences as channelization codes in the IS-95 standard [1]-[2] is only limited to its down-link channels that are synchronous. On the other hand, its up-link channels do not use Walsh-Hadamard sequences for channelization due to the fact that they perform extremely poor with very high aperiodic CCFs in asynchronous channels. Therefore, it is a totally different story here to talk about an isotropic MAI-free ISC and an orthogonal code family. To ensure an isotropic MAI-free operation a proper joint-selection of ISC and ISM is very important, such as the example of CC-CDMA system that has been shown in this sub-section and works on CC codes (as the ISC) and offset stacked spreading modulation (as the ISM).

5.4 MI-free IST and CC/DS-CDMA

Multipath interference (MI) is another serious impairing factor that the B3G wireless systems have to deal with, which is further complicated by the fact that extremely high data rate required in the B3G systems makes it more likely to happen that the transmitting signal bandwidth becomes much wider than channel coherent bandwidth, thus been suffered from severe frequency-selective fading. The traditional counter-measure against the frequency-selective fading is to use a RAKE

receiver, consisting of several parallel correlator *fingers* to capture different multipath returns and then combine them in either a coherent or non-coherent way to achieve multipath diversity. We have to admit that the RAKE does work well in most circumstances in traditional 2-3G systems to overcome the MI. However, our question here is that whether it is the *best* solution for the 2-3G as well as future B3G systems. The answer should be negative due to the following two main arguments. First, the capability for each finger of a RAKE receiver to separate or isolate individual multipath return is limited because, even if a finger could track the instance that a particular multipath return appears, the output from that finger is inevitably the combination of all possible unwanted interference from other users' transmissions plus the side lobes of other multipath returns of its self-transmission, in addition to the useful auto-correlation peak. Thus, the signal-to-interference ratio at the output of a RAKE finger is sensitively affected by the other unwanted transmissions and the side-lobes of other multipath returns of its self-transmission. Second, for a RAKE to work satisfactorily, a receiver needs to acquire virtually all information about the channel, i.e., the delay profile or impulse response of the channel, before a maximal ratio combining (MRC) RAKE algorithm can be performed. Therefore, the channel estimation is a must for any MRC-RAKE, adding a great complexity burden to overall receiver hardware. Therefore, it sounds much more desirable for a CDMA system to overcome the frequency-selective fading by its own signaling merits, rather than depending on a complex receiver, such as MRC-RAKE and etc. In other words, it will be strongly recommended that the signaling of a CDMA system could be implanted some inherent resistance against frequency-selective fading, which is what we want to get from the isotropic MI-free IST's. The very core of the IST's is the use of isotropic spreading codes with an ideal property that ensures all non-zero relative shifted versions of autocorrelation function to be zero in either synchronous or asynchronous channels, except for the autocorrelation peak at the zero shift. In doing so, delayed multipath returns will never cause any harmful interference to the precedent main-path signal detection when sampled at the instances of the main-path arrivals. Our very recent study has revealed that some types of complementary codes do exhibit such ideal isotropic MI-free property and they can possibly be implemented by using simple digital technology. With the help of those IST's, an isotropic MI-free operation environment can be created with or without RAKE receiver, greatly improving signal reception quality and reducing the receiver hardware complexity.

As already demonstrated in the previous subsection [69], the combination of CC codes and offset stacked spreading modulation can offer

an isotropic MAI-free operation. Unfortunately, it does not offer an isotropic MI-free operation due mainly to the use of the offset stacked spreading modulation, which causes strong mutual interference from different multipath returns. Nevertheless, it is very interesting to know from our recent study that the combination of the CC codes and conventional direct sequence spreading modulation can offer an isotropic MI-free operation, while still retaining its isotropic MAI-free property. With those desirable features, a CC/DS-CDMA architecture has been proposed by us in this chapter for the possible B3G wireless applications, which succeeds in getting rid of all those malefic MAI and MI interferences and transforming a CDMA system back to work in a noise-limited environment.

Figure (5.25) shows the conceptual block diagram of a K -user CC/DS-CDMA system with the MAI-free and MI-free IST using CC codes (as the ISC) and DS spreading modulation (as the ISM). It is noted that the CC/DS-CDMA system shown in Figure (5.25) differs substantially from the CC-CDMA system proposed in Figures (5.11) and (5.12) [66] in terms of complexity of the transmitters as well as the receivers. Especially, the spread modulated signal in the CC/DS-CDMA system is no longer multi-leveled and thus a simple BPSK modulation is sufficient here, resulting a much simplified transceiver hardware if compared with that of the CC-CDMA system [69], whose working principle is shown in Figures (5.11) and (5.12).

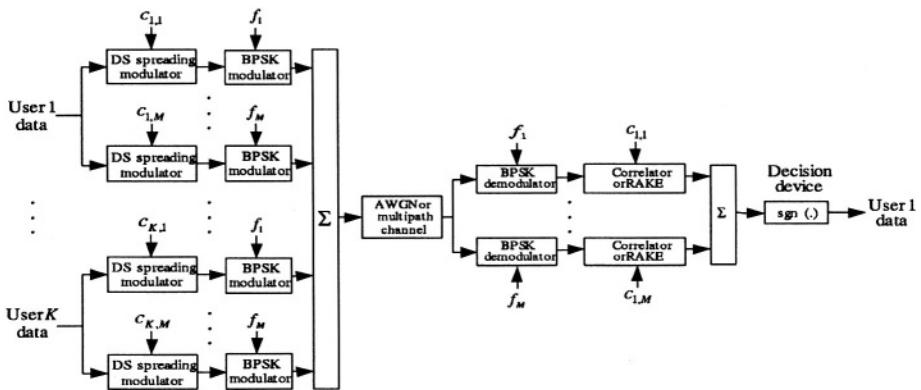


Figure 5.25. Block diagram of CC/DS-CDMA system with isotropic MAI-free and MI-free property, where the signal of interest to the receiver is user 1's transmission.

To illustrate the MI-free operation of the CC/DS-CDMA system shown in Figure (5.25), let us consider the following simple example of the complete complementary codes: $c_{11} = + + + -, c_{12} = + - + +, c_{21} = + - + +, c_{22} = + - - -,$ where we have assumed $K=M=2$ in this exam-

ple, corresponding to the parameters concerned in Figure (5.25). The users 1 and 2 are assigned c_{11} , c_{12} and c_{21} , c_{22} as their signature codes, respectively. Presumably the receiver wants to detect the signal from user 1, or c_{11} and c_{12} are the element codes of interest to the receiver. Figures (5.26) and (5.27) show the decoding process of the MI-free IST based on the complete complementary codes: c_{11} , c_{12} , c_{21} and c_{22} , as defined earlier. Without losing generality, it is further assumed that the user bit stream consists of three consecutive bits, whose values are $(b_{11}, b_{12}, b_{13}) = (+1, -1, +1)$ for user 1 and $(b_{21}, b_{22}, b_{23}) = (+1, +1, -1)$ for user 2 respectively, and there are two equal-strength multipath returns for each user's transmission, whose inter-path delay is one chip only. Figures (5.26) and (5.27) illustrate the detection process from the 1st bit to the 3rd bit in the up-link and down-link channels. In Figure (5.27) it is further assumed that the inter-user delay is also one chip for illustration clarity. The receiver, whose structure is shown in Figure (5.25) (note that RAKE is not used in this example), uses a simple matched filter or correlator to detect the signal from user 1. The decoding process proceeds bit by bit, given that the receiver has acquired bit synchronization with incoming signal. From Figures (5.26) and (5.27) it can be easily seen that the MI does not cause any impairing effect on the data decoding process, even though the receiver only uses a simple correlator. The values assumed for the inter-path delay, inter-user delay and number of multipath returns in Figures (5.26) and (5.27) are only for the illustration clarity and simplicity, and the same results will apply even for any other settings of inter-path and inter-user delays (different from one chip). The readers can verify it by drawing the similar graphs by their own. This interesting result tells us the fact that the combinatorial use of CC codes and traditional DS spreading can become the core of a new type of CDMA system, or CC/DS-CDMA architecture, which creates a truly isotropic MI-free operation environment with a simple correlator, without using any complex receiver implementations, such as MRC-RAKE, etc.

As pointed out earlier, the CC/DS-CDMA architecture shown in Figure (5.25) can also offer an isotropic MAI-free property, in addition to its isotropic MI-free property, as shown in Figures (5.28) and (5.29), where exactly the same CC codes, the DS spreading modulator, the information bit stream pattern and receiver structure as shown in Figures (5.26) and (5.27) are concerned. It is evident that the combined isotropic MAI-free and isotropic MI-free properties of the proposed CC/DS-CDMA architecture can offer a superior air-link performance for the new generation of wireless communications, which can finally make our dream come true: turning an interference-limited CDMA into a noise-limited system, leav-

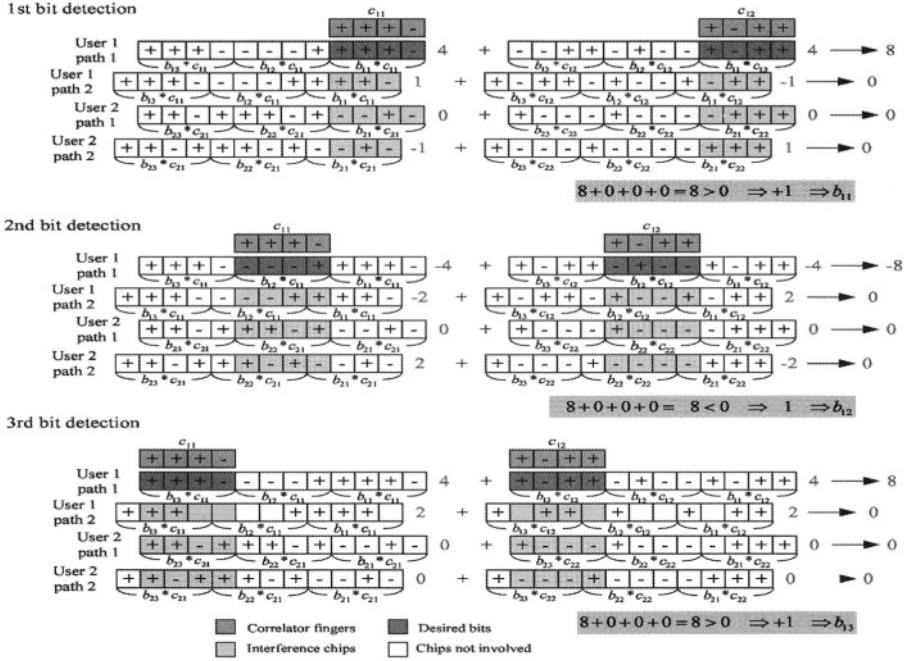


Figure 5.26. Illustration of MI-free operation for a 2-user CC/DS-CDMA system in synchronous down-link channels, where a two-path multipath channel is considered with the inter-path delay being one chip.

ing us a great room to push system capacity and data transmission rate into an even higher level than ever possible before.

It is not surprising that the proposed CC/DS-CDMA system can also employ a RAKE to further improve its detection efficiency, while retaining the isotropic MAI-free and isotropic MI-free properties. Figures (5.30) and (5.31), where a two-finger maximum ratio combining RAKE is considered in a 2-ray multipath channel with equal strength and only the detection procedure for the 2nd bit is shown for illustration clarity, are given to illustrate the detection process in up-link and down-link channels, respectively, in the same operational scenario as concerned in Figures (5.28) and (5.29). It is observed from Figures (5.30) and (5.31) that absolute value of the decision variable formed at the output of a MRC-RAKE in the receiver has been doubled (equal to 16) if compared to that (being 8) with a simple correlator in Figures (5.28) and (5.29). Such a result can be interpreted as an improved immunity against noise and other malicious external interference from the channel. Figures (5.32) to (5.34) are the BER performance to verify by sim-

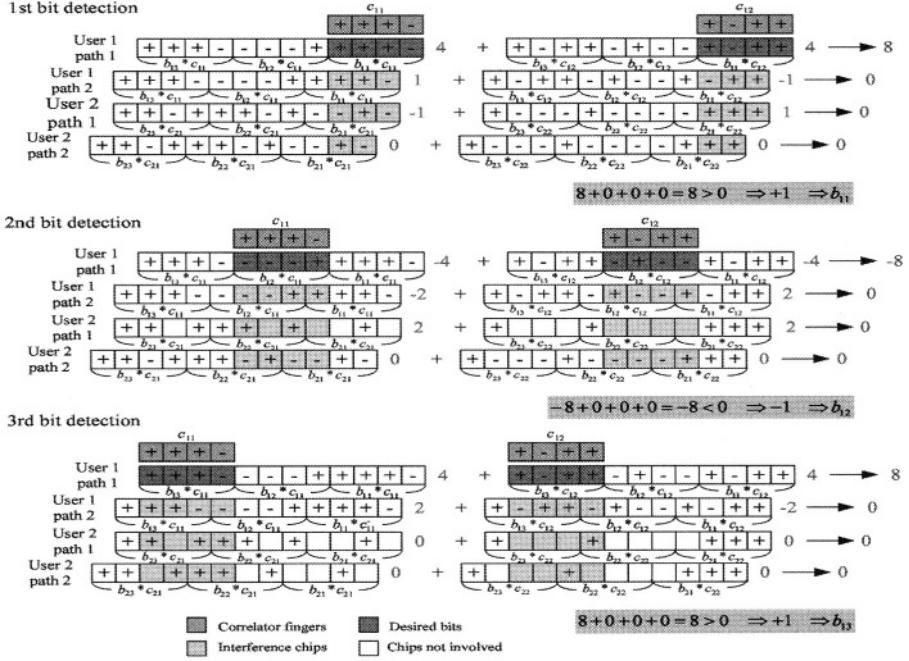


Figure 5.27. Illustration of MI-free operation for a 2-user CC/DS-CDMA system in asynchronous up-link channels, where a two-path multipath channel is considered with the inter-path delay being one chip and inter-user delay being also one chip.

ulations the three major characteristics features of the CC/DS-CDMA system: its MAI-free property, its ideal isotropic link performance and its inherent frequency-selective fading resistance or MI-free property, respectively. In particular, Figure (5.32) offers a comparison between the CC/DS-CDMA and conventional DS-CDMA with either Gold or Walsh-Hadamard codes in terms of their impairing effect of MAI on the BER performance. For each system, two scenarios are compared: one is a single user system and the other is the 4-user system. The former represents the operation without MAI and the latter studies the performance under MAI.

It is seen from Figure (5.32) that the CC/DS-CDMA system offers a perfect MAI-free operation, indicated by the almost overlapped two solid curves (with circle and cross markers) in the figures. Figure (5.33) compares the effect of link operation modes (either asynchronous up-link or synchronous down-link) on the BER performance of the CC/DS-CDMA and DS-CDMA systems. Again, the two solid curves (with circle and cross markers) are almost coincided, showing that the air-link opera-

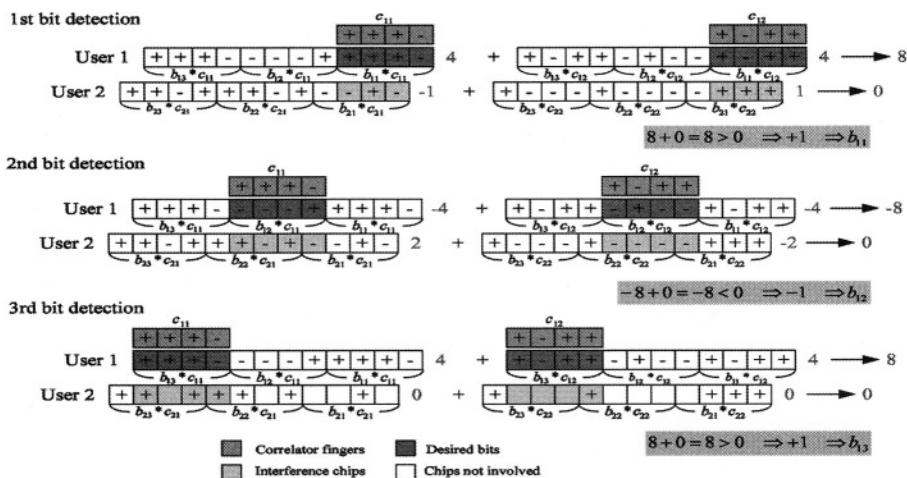
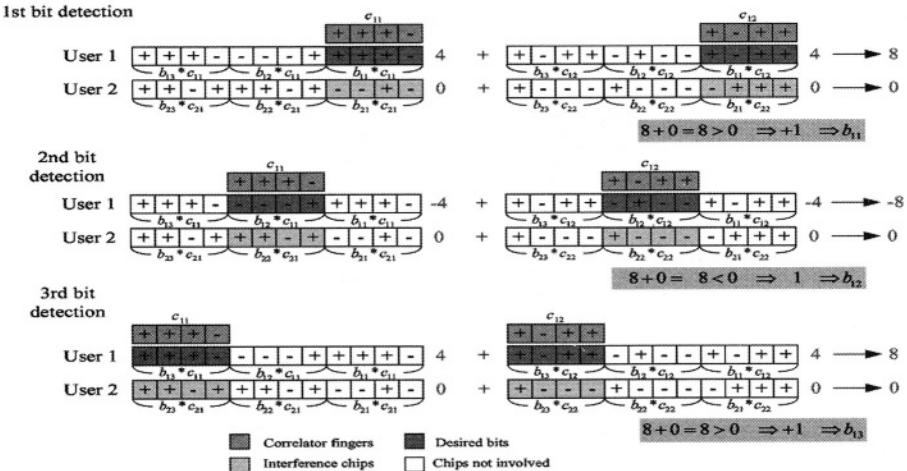


Figure 5.29. Illustration of MAI-free operation for a 2-user CC/DS-CDMA system in asynchronous up-link channels, where inter-user delay is assumed to be one chip.

tion mode exerts little impact on the overall link BER performance of a CC/DS-CDMA system. In other words, Figure (5.33) just verifies that a CC/DS-CDMA system indeed retains an isotropic operation property, which is independent of link operation modes. The frequency-selective fading resistance for different systems is shown in Figure (5.34), where the CC/DS-CDMA outperforms the other two counterparts with its al-

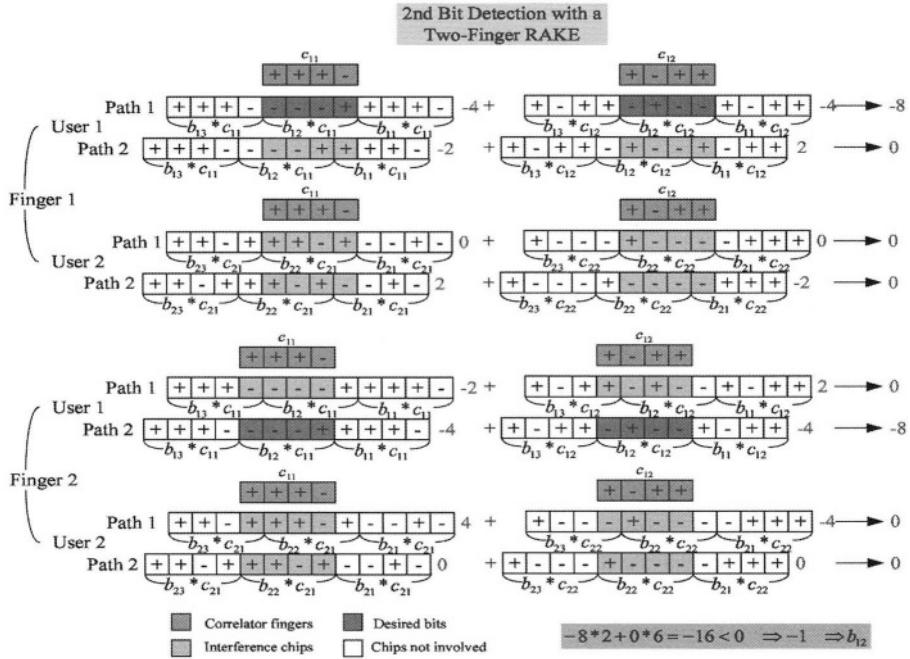


Figure 5.30. Illustration of RAKE receiver operation for a 2-user CC/DS-CDMA system in synchronous down-link channels where inter-path delay is assumed as one chip.

most identical performance under two very much different operational scenarios: the AWGN channel and 4-path multipath channel with each path having equal strength, representing a severe multipath interference or frequency-selective fading. We can also see from Figures (5.32) to (5.34) that the DS-CDMA system with 64-chip Walsh-Hadamard codes offers the worst performance, being extremely sensitive to MAI, link-operational modes and frequency-selective fading effects. The performance for the DS-CDMA system using 63-chip Gold code (indicated by two long-dash curves with circle and cross markers in Figures (5.32) to (5.34)) is in between the CC/DS-CDMA and Walsh-Hadamard DS-CDMA systems.

To show how well a system can perform relatively to a noise-limited system, the BER for a BPSK system in AWGN channel is also plotted in Figures (5.32) to (5.34), as given by a solid line curve (without marker). It can be observed from the figures that the CC/DS-CDMA system offers a performance very close to that for a BPSK system under AWGN channel, regardless of MAI, MI and the link operation modes (synchro-

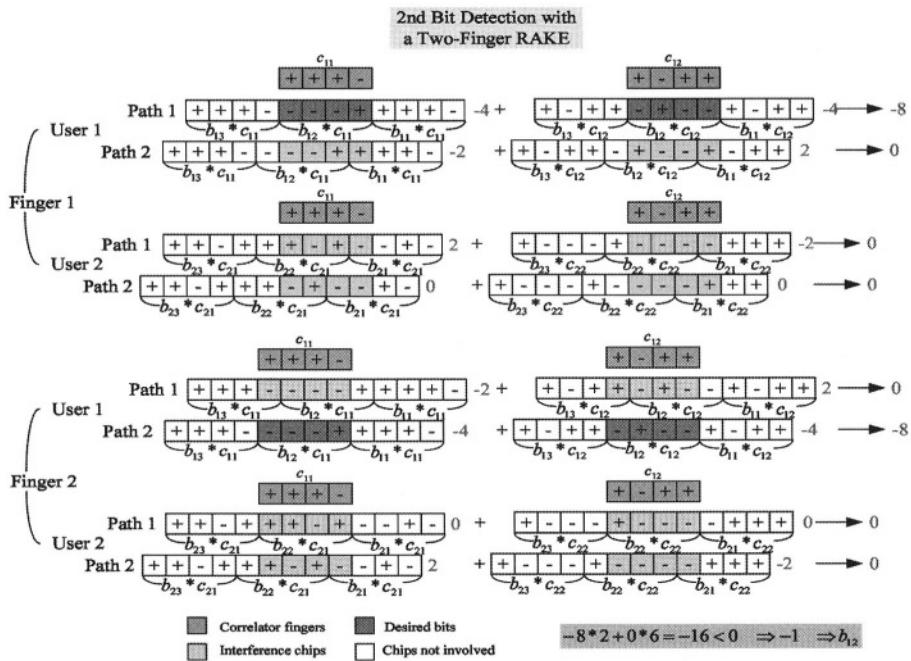


Figure 5.31. Illustration of RAKE receiver operation for a 2-user CC/DS-CDMA system in asynchronous up-link channels where both inter-path delay and inter-user delay are assumed to be one chip.

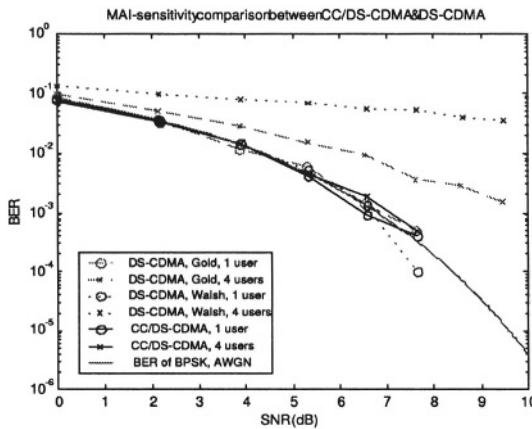


Figure 5.32. Comparison of conventional DS-CDMA and CC/DS-CDMA systems in terms of their MAI-resistance.

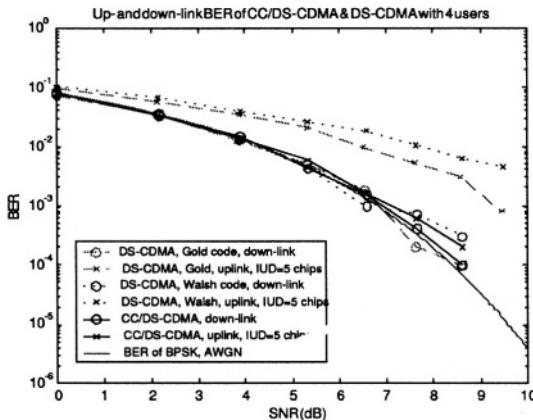


Figure 5.33. Comparison of conventional DS-CDMA and CC/DS-CDMA systems in terms of their sensitivity against operation modes (in either synchronous or asynchronous channels). IUD stands for *inter-user delay*.

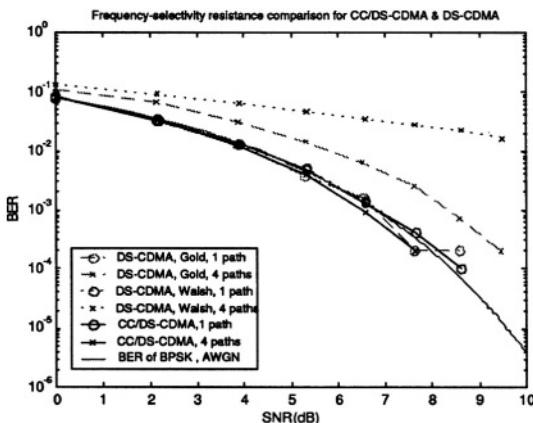


Figure 5.34. Comparison of conventional DS-CDMA and CC/DS-CDMA systems in terms of their ability to mitigate multipath interference or frequency-selective fading.

nous or asynchronous). This observation can be translated into that the performance of a CC/DS-CDMA system is indeed noise-limited. However, the performance of a conventional DS-CDMA system using either orthogonal codes (such as Walsh-Hadamard code) or quasi-orthogonal codes (such as Gold code) is clearly interference-limited.

5.5 Merits of CC/DS-CDMA

We have already demonstrated that the CC/DS-CDMA system architecture based on the IST can offer isotropic MAI-free and isotropic MI-free operation, which attribute to several operational advantages otherwise not possible with conventional CDMA technologies.

- First, it is not necessary for the CC/DS-CDMA system to use a RAKE receiver to mitigate frequency-selective fading in the channel. Only a simple correlator or matched filter is sufficient to yield satisfactory detection efficiency, thus greatly reducing the hardware cost of the receivers. Similarly, it also paves a way for a CC/DS-CDMA receiver to detect signals in a truly blind fashion without any prior channel information estimation. On the other hand, a RAKE has to acquire detailed channel information, such as delays and amplitudes of all multipath returns, for its successful operation of maximal-ratio-combining (MRC) algorithm, whose costly hardware in a mobile handset should never be underestimated.
- Second, as mentioned earlier, the CC/DS-CDMA receiver can also use a RAKE to separate different multipath returns to realize a genuine *multipath diversity* for further improved detection efficiency. It should be emphasized that the use of a RAKE in the CC/DS-CDMA system can yield a much better results than that possible in a conventional DS-CDMA system, because the output from each finger in the CC/DS-CDMA system consists of only autocorrelation peak at a particular time and nothing else. On the contrary, the output from each finger in a conventional CDMA system is a mixture of useful part (the autocorrelation peak at the instance) as well as a wealth of other unwanted components, including those nontrivial side lobes of other multipath returns in both self-transmission and other users' transmissions. Therefore, the CC/DS-CDMA system is in particular suitable for the operation in the channels with severe multipath interference due to its unique frequency-selective-fading resistance, as shown in Figure (5.34). Furthermore, if a RAKE is used in a CC/DS-CDMA receiver, the equal gain combining (EGC) can be considered to yield a satisfactory detection efficiency, due to the fact that the output from each finger in a CC/DS-CDMA receiver contains only the useful part: the autocorrelation peak at a particular sampling time. The advantage to use EGC instead of MRC in a CC/DS-CDMA RAKE is to save a complicated multipath amplitude estimation

algorithm. Thus, an EGC-RAKE in a CC/DS-CDMA receiver requires only multipath return delays information, which is much easier to obtain than the multipath return amplitudes.

- Third, as shown in Figure (5.32), the isotropic MAI-free property in the CC/DS-CDMA architecture makes it unnecessary to use multi-user detection to de-correlate the transmission signals from different users, due to the fact that the transmissions from different users in the CC/DS-CDMA system are already pre-decorrelated at transmitter side because of its MAI-free signaling structure. In other words, the signaling design in the CC/DS-CDMA architecture has already taken into account multi-user signal decorrelation mechanism, such that the signal detection efficiency in a CC/DS-CDMA receiver can be substantially improved even without using a complex and error-prone multi-user detection algorithm.
- Fourth, owing to the isotropic MAI-free and MI-free properties, the near-far effect will virtually cause no harm to the signal detection process at a correlator receiver in a CC/DS-CDMA system, as long as bit-synchronization, which should not be too difficult to implement with the help of a pilot, can be achieved prior to the data detection process. In other words, the CC/DS-CDMA is a deliberated-designed system with an excellent near-far resistance. Therefore, a complicated open-loop and closed-loop power control system is no longer a necessity to CC/DS-CDMA, saving a great amount of hardware and software overhead in user terminals. More precisely, the power control in the CC/DS-CDMA system is used merely to reduce power consumption and unnecessary power emission at terminals, whose requirements on response time and power control accuracy can be made much more relaxed than necessary in a conventional CDMA system.
- Fifth, the CC/DS-CDMA system in principle does not care about the number of co-channel unwanted transmissions within a cell because of its isotropic MAI-free property. Therefore, the signal processing advantage from beam forming provided by smart-antenna technology has become less relevant to the improvement in signal detection efficiency of a CC/DS-CDMA system, as long as the system does not need users' particular spatial information, such as direction-of-arrivals of the incoming signals. More specifically, the beam-patterns adopted at a CC/DS-CDMA receiver have little direct impact on the signal-to-interference-cum-noise ratio if compared to that in a conventional CDMA system without isotropic

MAI-free and MI-free properties. Therefore, an omni-directional antenna is enough for CC/DS-CDMA system to perform satisfactorily, without needing costly antenna array system.

- Finally, the CC/DS-CDMA scheme is in particular well suited for detection of the data appearing at the edges of a message, as shown in the upper parts of Figures (5.26), (5.27) and (5.29), which illustrate the signal detection process for the first (or leftmost) bit of the message. In detecting those bits, partial CCFs of the codes will play an important role. It is seen from the figures that the CC/DS-CDMA system yields zero partial CCFs, which ensure an ideal performance for signal detection even on the edges of a message or packet. This observation is significant due to the concern of dominant bursty-type traffic in future wireless systems, where frequent on-and-off operating fashion in packet-switched applications will force a receiver to work constantly on detection of packet-edges. Therefore, in this sense, the CC/DS-CDMA is a perfect choice for the future all-IP wireless applications where bursty traffic will be dominant.

In summary, the CC/DS-CDMA architecture based on IST's can create an interference-free operational environment due to its superior isotropic MAI-free and isotropic MI-free properties, which can successfully transform an interference-limited operational environment into a noise-limited one. Thus, a great gain in the capacity and data throughput in a CDMA system can be expected. The major operational advantages of the CC/DS-CDMA architecture, if compared with a conventional DS-CDMA, are listed in Table (5.6).

6. Summary

CDMA technology is at a critical stage on its evolutional path. The current major CDMA-based standards all work on almost the same principle with direct-sequence spreading modulation and conventional spreading codes, suffering from serious interference problems stemmed from their undesirable spreading codes characteristics, and thus their performance is strictly limited by combined effect of extremely unpredictable MAI and MI. The conventional spreading codes applied to various CDMA-based 2-3G standards are never optimal in terms of their correlation properties as well as their asymmetrical feature when working in synchronous down-link and asynchronous up-link, causing many problems in rate-matching, channelization and co-channel inter-

Table 5.6. Comparison between CC/DS-CDMA and conventional DS-CDMA.

| | CC/DS-CDMA | Conventional DS-CDMA |
|-----------------------------|---------------|-------------------------|
| Detection Efficiency | Excellent | Medium~good (with RAKE) |
| Multipath Diversity | Excellent | Medium |
| RAKE Receiver | Optional | A Must |
| FSF ^a Resistance | Excellent | None (without RAKE) |
| Near-far Resistance | Excellent | None |
| Precision Power Control | Not Necessary | A Must |
| Multi-User Detection | Not Necessary | Recommended |
| Up-Link Sync. Control | Not Necessary | Recommended |
| Smart Antenna | Not Necessary | Recommended |
| Rx Hardware | Simple | Complex |
| System Capacity | High | Low~medium |
| Down-Link MAI | Low | Low |
| Up-Link MAI | Low | Medium~high |
| Down-Link MI | Low | Medium~high |
| Up-Link MI | Low | Medium~high |
| Bursty traffic | Well-suited | Poorly-performed |

^a FSF: Frequency-Selective Fading.

ference, etc. The concept of isotropic spreading technologies suggested in this chapter serves well as a direction finder for the future research on CDMA technological development, based on which two innovative CDMA architectures, CC-CDMA and CC/DS-CDMA, are introduced herein. In particular, the CC/DS-CDMA system can offer a superior isotropic interference-free operation in a wireless communication system, showing a great potential for CDMA systems to offer a real noise-limited performance. We have a strong belief that the B3G wireless should definitely work on a completely new CDMA platform based on interference-free architecture, rather than on the conventional CDMA technologies, which have suffered from irreducible interference. The bandwidth-efficiency of a B3G CDMA system should be achieved only based on its superior power-efficiency, which is the bottleneck to hinder any further improvement on overall performance of current 2-3G wireless. Isotropic interference-free CDMA technologies, which can greatly improve the power-efficiency and then the overall performance, should play an indispensable role in the future wireless air-link architecture. The proposed two novel CDMA schemes have shown the way leading to the development of a practical noise-limited CDMA architecture, which is our ultimate goal to achieve on CDMA evolution.

Acknowledgments

The author wishes to thank the Editor of this book, Prof. Mohsen Guizani, for his encouragement and support in preparing this chapter. He would like also to thank Mr. Jin-Xiao Lin, Mr. Hsin-Wei Chiu and Mr. Zhemin Lin, Institute of Communications Engineering, National Sun Yat-Sen University, Taiwan, for their help in generating some figures and plottings presented in this chapter. The author would like also to thank National Science Council (NSC), Taiwan, for the following research grants, NSC 89-2213-E-005-032, NSC 89-2213-E-005-029, NSC 89-2213-E-005-032, NSC 90-2213-E-110-054, NSC 91-2213-E-110-032, NSC 92-2213-E-110-015, NSC 93-2213-E-110-012, NSC 92-2213-E-110-047 and NSC 92-2213-E-110-048, which have partially supported the related research work.

Appendix: Complete Complementary Codes with PG being 8 to 512

The following list gives the complementary codes with their PG values changing from 8 to 512, which are commonly used in a practical CDMA system. Note that the commons “,” are used to separate different element codes within a specific flock and blankets “()” are used to indicate flocks of a complementary code sets.

1 PG=8

Length of element codes=4, flock size=2

(+ + + -, + - + +)

(+ + - +, + - - -)

2 PG=16

Length of element codes=8, flock size=2

(+ + + - + + - +, + - + + + + - - -)

(+ + + - - + -, + - + + - + + +)

3 PG=32

Length of element codes=16, flock size=2

(+ + + - + + - + + + + - - + -, + - + + + - - + - + + - + + +)

$$(+ + + - + + - + - - + + + - +, + - + + + - - - + - + - -)$$

4 PG=64

(1) Length of element codes=32, flock size=2

(2) Length of element codes=16, flock size=4

$(+++++-+--++-+-+, +--+---++++)$

(+ + + + + - - - + + - + + -, + - + - + + + + - + + - - - + +, +

(+ + + + - + - + + + - +, + - + - - - + + - + + -, + +

-- + + - - + - -, + - + - + + - - + + + + +)
PG_128

Language

Length of element codes = 4; block size = 2

6 PG=256

(1) Length of element codes=128, flock size=2

(2) Length of element codes=64, flock size=4

7 PG=512

(1) Length of element codes=256, flock size=2

(2) Length of element codes=64, flock size=8

- + + + + + + + - + + - + - + + - - + + + - + + - + - + - + +
- - - + + + +, + + + + - - - + - + + - + - - + + + + - + - -
+ - + + - - - - + - + - + - + + - + + - + + - + + - + + - + +
-, + - + - + - + - - + + + + - + + - + - + + + - - + + - + +
+ - + - + + + + + + + + - + + - + - + + + + - + + - + + - + +
- + + - + + + - - + - - + + + + + - + - + - + + - + + - + + - +
+ - - + + - - + + + + + + + + - + - + - + +, + - - + - + + - -
+ + + + - - + - + + - + + + + + - - + + - + - + + - + + - + - +
+ - - + - + - + - + - - -)

References

- [1] TIA/EIA-95A, (1995). Mobile Station-Base Station Compatibility Standard for Dual-Mode Wideband Spread Spectrum Cellular Systems.
- [2] TIA/EIA-95B, (1997). Mobile Station-Base Station Compatibility Standard for Dual-Mode Wideband Spread Spectrum Cellular Systems. Baseline Version, July 31, 1997.
- [3] Telecommunications Industry Association (TIA), (1998). The cdma2000 ITU-R RTT Candidate Submission (0.18).
- [4] ETSI, (1998). The ETSI UMTS Terrestrial Radio Access (UTRA) ITU-R RTT Candidate Submission, Jan. 29, 1998.
- [5] Association of Radio Industries and Businesses (ARIB), (1998). Japan's Proposal for Candidate Radio Transmission Technology on IMT-2000: W-CDMA, June 26, 1998.
- [6] CATT, (1998). TD-SCDMA Radio Transmission Technology For IMT-2000 Candidate submission, Draft V.0.4, Sept. 1998.
- [7] (2002). First Report and Order in the Matter of Revision of Part 15 of the Commission's Rules Regarding Ultra-Wideband Transmission Systems. FCC, released, ET Docket 98-153, FCC 02-48, April 2002.
- [8] Mitchell T. (2001). Broad is the way. *IEE Review*, pp. 35-39, Jan. 2001.
- [9] Leeper D. (2001). A long-term view of short-range wireless. *IEEE Computer*; 34(6): pp. 39-44, 2001.
- [10] Van't Hof JP, Stancil DD. (2002). Ultra-wideband high data rate short range wireless links, Pages: 85-89, Proc. IEEE VTC 2002.
- [11] Win MZ. (2002). A unified spectral analysis of generalized time-hopping spread-spectrum signals in the presence of timing jitter. *IEEE J. Select. Areas Communications*; 20(9): 1664-1676, 2002.
- [12] Win MZ, Scholtz RA. (2002). Characterization of ultra-wide bandwidth wireless indoor channels: a communication-theoretic view. *IEEE J. Select. Areas Communications*, 20(9): 1613-1627, 2002.
- [13] United States Code of Federal Regulations, Title 47, Part 15, U.S. Government Printing Office, <http://www.access.gpo.gov/nara/cfr/index.html>.
- [14] Conroy JT, LoCicero JL, Ucci DR. (1999). Communication techniques using monopulse waveforms. Proc. MILCOM 1999: 2:1185-1191.
- [15] Ross, A. H. M., and K. L. Gilhousen, (1996). CDMA Technology and the IS-95 North American Standard, in The Mobile Communications Handpaper, J. D. Gipson (Ed.), CRC Press, pp. 430-448, 1996.

- [16] N. Guo and L. B. Milstein, (1999). On rate-variable multidimensional DS/SSMA with dynamic sequence sharing, *IEEE J. Select. Areas Commun.*, vol. 17, May 1999.
- [17] C.-L. I and R. D. Gitlin, (1995). Multi-code CDMA wireless personal communications networks, in Proc. IEEE Int. Conf. Commun. (ICC'95), Seattle, WA, vol. 2, pp. 1060-1064.
- [18] S. Sasaki, H. Kikuchi, H. Watanabe, and J. Zhu, (1994). Performance evaluation of parallel combinatory SSMA systems in Rayleigh fading channel, in Proc. IEEE 3rd Int. Symp. Spread Spectrum Techniques and Applications (ISSSTA'94), Oulu, Finland, vol. 1, pp. 198-202.
- [19] Baey, S.; Dumas, M.; Dumas, M.-C., (2002). QoS tuning and resource sharing for UMTS WCDMA multiservice mobile, *IEEE Trans. on Mobile Computing*, Vol. 1, Issue: 3 , Page(s): 221 -235, Jul-Sep 2002.
- [20] Insoo Sohn; Seung Chan Bang; (2000). Performance studies of rate matching for WCDMA mobile receiver, Vehicular Technology Conference, 2000. IEEE VTS-Fall VTC 2000. 52nd , Vol. 6 , Page(s): 2661 -2665, 2000.
- [21] Kam, A.C.; Minn, T.; Siu, K.-Y. (2001). Supporting rate guarantee and fair access for bursty data traffic in W-CDMA, *IEEE J. Select. Areas Communications*, Volume: 19, Issue: 11, Page(s): 2121-2130, Nov 2001.
- [22] Thit Minn; Kai-Yeung Siu, (2000). Dynamic assignment of orthogonal variable-spreading-factor codes in W-CDMA, *IEEE J. Select. Areas Communications*, Vol. 18, Issue: 8, Page(s): 1429-1440, Aug 2000.
- [23] Fantacci, R.; Nannicini, S., (2000). Multiple access protocol for integration of variable bit rate multimedia traffic in UMTS/IMT-2000 based on wideband CDMA, *IEEE J. Select. Areas Communications*, Vol. 18, Issue: 8, Page(s): 1441-1454, Aug. 2000.
- [24] R. Gold, (1968). Maximal recursive sequences with 3-valued recursive cross-correlation functions, *IEEE Trans. on Information Theory*, vol. IT-14, pp. 154-156, January 1968.
- [25] R. A Scholtz and L. R. Welch. (1984). GMW Sequences. *IEEE Trans. on Information Theory*, Vol. 30, No. 3, pp. 548-553, May 1984.
- [26] J-S. No and P. V. Kumar. (1989). A New Family of Binary Pseudorandom Sequences Having Optimal Periodic Correlation Properties and Large Linear Span. *IEEE Trans. on Information Theory*, Vol. 35, No. 2, pp. 371-379, March 1989.

- [27] X. H. Chen and J. Oksman, (1992). BER performance analysis of 4-CCL and 5-CCL codes in slotted indoor CDMA systems, *IEE Proceedings - I*, vol. 139, pp. 79-84, February 1992.
- [28] X. H. Chen; Lang, T.; Oksman, J.; (1996). Searching for quasi-optimal subfamilies of m-sequences for CDMA systems, Seventh IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC'96), vol. 1, Page(s): 113 -117, 15-18 Oct. 1996.
- [29] Tirkel, A.Z. (1996). Cross correlation of m-sequences-some unusual coincidences, Spread Spectrum Techniques and Applications Proceedings, 1996 IEEE 4th International Symposium on, Page(s): 969 -973 vol.3, 22-25 Sept. 1996.
- [30] Ito, T.; Sampei, S.; Morinaga, N., (2000). M-sequence based M-ary/SS scheme for high bit rate transmission in DS/CDMA systems, *Electronics Letters*, vol. 36 Issue: 6, Page(s): 574 -576, 16 Mar 2000.
- [31] Tirkel, A.Z., (1996). Cross correlation of m-sequences-some unusual coincidences, Spread Spectrum Techniques and Applications Proceedings, 1996 IEEE 4th International Symposium on, vol. 3, Page(s): 969 -973, 22-25 Sep 1996.
- [32] Imamura, K.; Guo-Zhen Xiao, (1992). On periodic sequences of the maximum linear complexity and M-sequences, Singapore ICCS/ISITA '92. 'Communications on the Move' , Page(s): 1219 -1221 vol.3, 16-20 Nov. 1992.
- [33] Uehara, S.; Imamura, K., (1992). Some properties of the partial correlation of M-sequences, Singapore ICCS/ISITA '92. 'Communications on the Move', Page(s): 1222 -1223, vol.3, 16-20 Nov. 1992.
- [34] Turkmani, A.M.D.; Goni, U.S., (1993). Performance evaluation of maximal-length, Gold and Kasami codes as spreading sequences in CDMA systems, Universal Personal Communications, 1993. Personal Communications: Gateway to the 21st Century Conference Record, 2nd International Conference on, vol. 2, Page(s): 970 -974, 12-15 Oct. 1993.
- [35] Lahtonen, J, (1995). On the odd and the aperiodic correlation properties of the Kasami sequences, *IEEE Trans. on Information Theory*, vol. 41 Issue: 5, Page(s): 1506-1508, Sept. 1995.
- [36] Lebedev, O.N.; Poliakov, I.L., (2000). Properties of composite Kasami sequence sets for wideband signals, 10th International Microwave Conference, 2000. *Microwave and Telecommunication Technology*, Page(s): 234 -235, 2000.
- [37] Barghouthi, R.T.; Stuber, G.L.; (1994). Rapid sequence acquisition for DS/CDMA systems employing Kasami sequences, *IEEE*

- Trans. on Communications*, vol. 42, Issue: 2, Page(s): 1957-1968, Feb/Mar/Apr 1994.
- [38] Komo, J.J.; Liu, S.-C.; (1990). Modified Kasami sequences for CDMA System Theory, Twenty-Second Southeastern Symposium on , Page(s): 219 -222, 11-13 Mar. 1990.
- [39] Dongwook Lee; Hun Lee; Milstein, K.B.; (1998). Direct sequence spread spectrum Walsh-QPSK modulation, *IEEE Trans. on Communications*, vol. 46 Issue: 9, Page(s):1227 -1232, Sept. 1998.
- [40] Joonyoung Cho; Youhan Kim; Kyungwhoon Cheun; (2000). A novel FHSS multiple-access network using M-mary orthogonal Walsh modulation, VTC 2000. IEEE VTS-Fall VTC 2000. 52nd, vol. 3, Page(s): 1134 -1141, 2000.
- [41] Tsai, S.; Khaleghi, F.; Seong-Jun Oh; Vanghi, V.; (2001). Allocation of Walsh codes and quasi-orthogonal functions in cdma2000 forward link, VTC 2001 Fall. IEEE VTS 54th, Page(s): 747 -751 vol.2, 2001.
- [42] M. J. E. Golay, (1961). Complementary series, *IRE Trans. Inform. Theory*, vol. IT-7, pp. 82-87, 1961.
- [43] R. Turyn, (1963). Ambiguity function of complementary sequences, *IEEE Trans. on Information Theory*, vol. IT-9, pp. 46-47, Jan, 1963.
- [44] N. Suehiro, (1982). Complete complementary code composed of N-multiple-shift orthogonal sequences, *Trans. IECE of Japan* (in Japanese), vol. J65-A, pp. 1247-1253, Dec. 1982.
- [45] N. Suehiro and M. Hatori, (1988). N-Shift Cross-Orthogonal Sequences, *IEEE Trans. on Information Theory*, vol. IT-34, no. 1, pp. 143-146, January 1988.
- [46] P. V. Kumar and R. A. Scholtz, (1983). Bounds on the linear span of Bent sequences, *IEEE Trans. on Information Theory*, pp. 854-862.
- [47] Tao Lang; Xiao-Hua Chen, (1994). Comparison of correlation parameters of binary codes for DS/CDMA systems, Singapore ICCS '94. Conference Proceedings, vol. 3, Page(s): 1059 -1063, 14-18 Nov. 1994.
- [48] L. R. Welch, (1974). Lower bounds on the maximum cross-correlation of signals, *IEEE Trans. on Information Theory*, vol. IT-20, pp. 397-399, 1974.
- [49] M. B. Pursley, (1977). Performance evaluation for phase-coded spread-spectrum multiple-access communications - Part I: System analysis, *IEEE Trans. on Communications*, vol. COM-25, no. 8, pp. 795-799, Aug. 1977.
- [50] M. B. Pursley and D. V. Sarwate, (1977). Performance evaluation for phase-coded spread-spectrum multiple-access communications - Part II: code sequence analysis, *IEEE Trans. on Communications*, vol. COM-25, no. 8, pp. 800-803, Aug. 1977.

- [51] K. Yao, (1977). Error probability of asynchronous spread spectrum multiple access communication systems, *IEEE Trans. on Communications*, vol. 25, no. 8, pp. 803-809, Aug. 1977.
- [52] D. V. Sarwate and M. B. Pursley, (1980). Cross-correlation properties of pseudorandom and related sequences, *Proceedings of the IEEE*, vol. 68, no. 5, pp. 593-620, May 1980.
- [53] M. B. Pursley, D. V. Sarwate and W. E. Stark, (1982). Error probability for direct-sequence spread spectrum multiple-access communications - Part I: Upper and lower bounds, *IEEE Trans. on Communications*, vol. COM-30, no. 5, pp. 975-984, May 1982.
- [54] E. A. Geraniotis and M. B. Pursley, (1982). Error probability for direct-sequence spread spectrum multiple-access communications - Part II: Approximations, *IEEE Trans. on Communications*, vol. COM-30, no. 5, pp. 985-995, May 1982.
- [55] J. M. Holtzman, (1992). On calculating DS/SSMA error probabilities, Proceedings of IEEE 2nd International Symposium on Spread Spectrum techniques and Applications (ISSSTA' 92), Yokohama, Japan, pp. 23-26, Dec. 1992.
- [56] CWTS WG1 LAS-CDMA, (2001). Physical channels and mapping of transport channels onto physical channels, LAS TS 25.221, V1.0.0, July 17-17, 2001.
- [57] Verdu, (1986). Minimum Probability of Error for Asynchronous Gaussian Multiple Access, *IEEE Trans. on Information Theory*, Vol. IT-32, No. 1, pp. 85-96, Jan. 1986.
- [58] Verdu, S., (1994). Adaptive Multiuser Detection, in Code Division Multiple Access Communications, S. G. Glisic and P. A. Leppanen, Eds., Boston: *Kluwer Academic Publishers*, pp. 97-116, 1994.
- [59] Moshavi, S., (1996). Multiuser Detection for DS-CDMA Communications, *IEEE Communications Magazine*, pp. 12-36, Oct. 1996.
- [60] Duel-Hallen, A., J. Holtzman, and Z. Zvonar, (1995). Multiuser Detection for CDMA Systems, *IEEE Personal Commun.*, pp. 4-8, April 1995.
- [61] Hsiao-Hwa Chen & H. K. Sim, (2001). A new CDMA multiuser detector - orthogonal decision-feedback detector for asynchronous CDMA systems, *IEEE Trans. on Communications*, vol. 49, no. 9, pp.1649-1658, Sept. 2001.
- [62] Hsiao-Hwa & Zhe-Qiang Liu, (2001). Zero-insertion adaptive minimum mean square error (MMSE) receiver for asynchronous CDMA multiuser detection, *IEEE Trans. on Vehicular Technology*, vol. 50, no. 2, pp. 557-569, March 2001.
- [63] Hsiao-Hwa Chen, Yi-Ning Chang & Yu-Bing Wu, (2003). Single Code Cyclic Shift Detection - A Pilot Aided CDMA Multiuser De-

- tector Without Using Explicit Knowledge of Signature Codes, *IEICE Trans. On Communications*, vol. E86-B, No. 4, pp. 1286-1296, April 2003.
- [64] R. Luples, S. Verdu, (1989). Linear Multi-user Detectors for Synchronous Code-Division Multiple-Access Channels, *IEEE Trans. On information Theory*, vol. 35, pp. 123-136, Jan 1989.
- [65] R. Luples, S. Verdu, (1990). Near-far Resistance of Multi-user Detectors in Asynchronous Channels, *IEEE Trans. on Communications*, vol. 38, pp. 496-508, April 1990.
- [66] Hsiao-Hwa Chen, T. Lang & J. Oksman, (1996). Correlation statistics distribution convolution (CSDC) algorithm for studying CDMA indoor wireless systems with RAKE receiver, power control & multipath fading," *IEICE Trans. on Communications*, vol. E79-B, No. 10, Oct. 1996.
- [67] CWTS TSM 01.00 V.2.0.0 (2001-07), China wireless Telecommunication Standard, 3G digital cellular telecommunications, TD-SCDMA System for Mobile (TSM), General Description, 2001.
- [68] Hsiao-Hwa Chen, CX Fan & Willie W. Lu, (2002) China's Perspectives on 3G Mobile Communications & Beyond:TD-SCDMA Technology, *IEEE Wireless Communications*, pp. 48-59, April, 2002.
- [69] Hsiao-Hwa Chen, J. F. Yeh & N. Seuhiro, (2001). A Multi-Carrier CDMA Architecture Based on Orthogonal Complementary Codes for New Generations of Wideband Wireless Communications, *IEEE Communications Magazine*, vol. 39, no. 10, pp. 126-135, October 2001.

This page intentionally left blank

Chapter 6

POWER CONTROL IMPLEMENTATION IN 3RD GENERATION CDMA SYSTEMS

Bassam Hashem

Research and Development Department
Saudi Telecom Company
Riyadh 11342
P.O.Box 261323
Saudi Arabia

Abstract Power control is one of the main features in 3rd generation cellular CDMA systems. Power control reduces the interference in the system and hence increases the capacity. In this paper, we describe the basic operation of power control implementation in 3G CDMA systems. The inner loop and the outer loop algorithms are explained where the power commands are sent at a high rate and uncoded resulting in high error rate in receiving them. The problem of errors in the power control commands becomes worse during soft handoff where the terminal communicates with more than one base station simultaneously. Few algorithms are introduced during soft handoff to increase the system capacity like power control rate reduction and the adjustment loop. Adjusting the transmission rate is discussed also as another scheme to achieve the required quality.

Keywords: CDMA, Power control, Soft handoff, adaptive modulation, rate control, UMTS, CDMA2000

6.1 Introduction

CDMA systems are characterized as being interference limited. Reducing the interference results in a direct increase in the system capacity. The uplink of CDMA cellular systems suffers from the near-far problem. If all users send a fixed power level, then the signal received from a nearby user will be stronger than a signal received from a far user and may mask it. This is

referred to as the near-far problem. Typically, the down link transmissions from the base station (BTS) to the mobiles employ orthogonal codes and hence the interference is reduced. Some of this orthogonality, however, is lost due to the multipath fading. This is in addition to the interference from other base stations in the downlink. Power control hence is employed on both the down and the up links.

Third generation systems are designed to serve terminals or UEs (users equipments) with different quality of service (QoS) requirements. Different QoS results in different block error rate (BLER) requirements which is governed by the energy per bit to noise ratio (E_b/N_0) where N_0 of course is both the interference and the white noise. Hence the objective of the power control algorithm is to make sure that each terminal is achieving its required E_b/N_0 . A lower E_b/N_0 results in higher BLER while a higher E_b/N_0 results in excessive interference which reduces the system capacity. It should be noted that it is not only the transmitted power that controls the achieved E_b/N_0 . The transmission rate and the interference level also affect the achieved E_b/N_0 . The rest of this chapter is organized as follows. Sections 6.2 to 6.6 covers the inner loop algorithm while section 6.7 covers the outer loop. Then we discuss the operation of power control during soft handoff (SHO) where section 6.8 covers the uplink and section 6.9 covers the downlink. Section 6.12 covers the operation of power control algorithms during compressed mode. Finally we discuss how the rate can be changed to achieve the required block error rate (BLER) and explain how adaptive modulation is investigated for future systems as a mean to replace power control.

6.2 Inner Loop Power Control

The received power varies depending on the distance between the BTS and the mobile, the shadowing and the multipath fading. Variations due to distance and shadowing are location dependent and hence are reciprocal on both the downlink and the uplink. Thus, the mobile can determine these variations by measuring the average received power on the downlink. This is called open loop power control. Full duplex is usually achieved by providing different frequency channels on the downlink and the uplink. The frequency separation between these channels is much larger than the coherence bandwidth of the communication channel and thus both links tend to fade independently. To eliminate the variations due to multipath fading, the base station sends power control commands to the mobile asking the mobile to either increase or decrease its power by a fixed amount. This is called closed

loop power control. Both the open loop and the closed loop constitute what is called the inner loop power control.

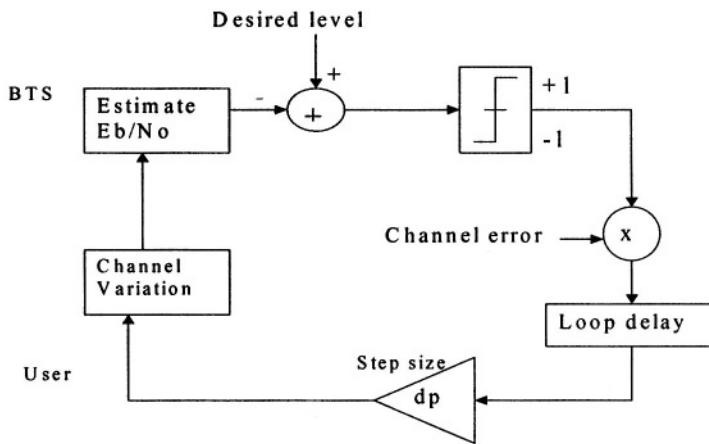


Figure 1 Inner loop power control

Figure 1 shows how the inner loop is implemented in 3G systems. This shows how the power control algorithm is implemented at the BTS to decide if the UE is to decrease or increase its power.

As was mentioned earlier, a similar algorithm is implemented at the UE to decide whether the BTS should increase or decrease its power. The base station will estimate the achieved Eb/No and compares it with the target Eb/No . The target Eb/No itself is controlled by the outer loop power control which will be explained later. If the achieved Eb/No is lower (higher) than the target Eb/No , the BTS sends an up (down) command asking the UE to increase (decrease) its power. This is done every slot or what is called sometimes power control group. The up command is sent as a 1 while the down command is sent as a 0 or vice versa (depending on the standard). The command it self may be received in error causing the UE to adjust its power in the wrong direction. Of course, the command will not reach the UE and be executed instantaneously but rather after some delay which is usually a one or two power control groups. After that, the UE adjusts the power by a certain step size. Standards support multiple step sizes. The received Eb/No can still not equal the target Eb/No due to variations in the channel conditions or the interference level. As we said earlier, in this paper, we assume No to include both the white noise and the interference.

6.3 Power control command rate

The variation in the communication channel is an essential factor in designing the power control algorithm. Such variation is a function of the

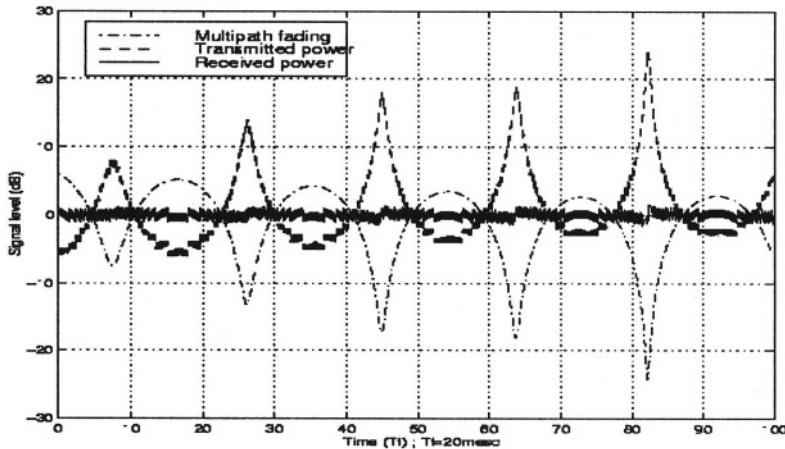


Figure 2 Variations in the channel condition

type of the communication channel (usually modeled as a Raleigh fading channel with multiple resolvable paths), the velocity of the UE and the carrier frequency. The rate at which the power commands is sent is determined by the rate at which the channel can change. Two common rates are 800 commands per second, which is adopted in the CDMA2000 standard, and 1500 commands per second, which is adopted in the UMTS (3GPP) standard. At low mobility, power control eliminates the variations in the received signal level and the received power can be made constant. This is shown in Figure 2 where 800 commands are sent to eliminate the variations in the channel condition for a UE moving at a speed of 5 Km/h in a two equal paths Raleigh fading channel. It is clear from the figure that the variations in the channel conditions can be eliminated at such a rate. This is usually referred to as perfect power control.

However, for fast users, power control will be slow to eliminate the variations in the received signal level completely and the residual error in the received signal power is known to follow a log-normal distribution with a standard deviation that is a function of the user velocity [3]. Many papers analyzing CDMA systems capacity do not include the impact of multipath fading on the system by assuming the power control algorithm to be perfect. This is, however, is only correct if the system has only one cell! A UE going into a deep fade transmits at high power level to compensate for the fade.

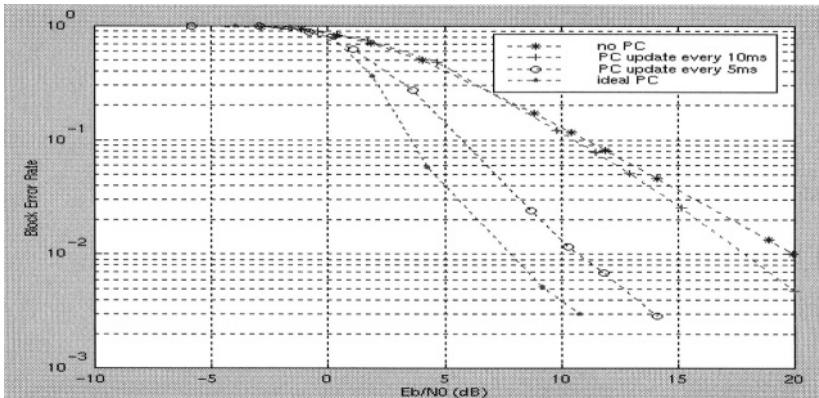


Figure 3: Performance comparison of raw BER vs. Eb/N0

This does not of course increases the intracell interference (interference from users communicating with the same BTS) and hence the impact of multipath fading in TDD can be neglected when estimating the intracell interference. However, transmitting at high power levels increases the intercell interference. Hence, variations in the channel conditions due to multipath fading may not be omitted in analyzing the multi cell system even if the power control is capable of tracking multipath fading. It is even shown that tracking deep fades reduces the system capacity due to the excessive intercell interference[3]. The ability to track deep fades is controlled by the dynamic range of the closed loop power control. A large dynamic range allows such tracking. A system that relies only on the closed loop (no open loop) power control to track variation in the channel due to distance, shadowing and multipath fading requires large dynamic range. The impact of such large dynamic range on increasing the intercell interference should be investigated carefully. Even though the open loop power control tracks changes due to distance and shadowing in an FDD system, the open loop suffers from a large estimation error and hence some designers may want to rely on the closed loop only [4].

The discussion on the difference capabilities of the open loop and the closed loop algorithm leads us to wonder if power control is needed in TDD mode. In TDD mode, both the transmitter and the receiver uses the same frequency band. Hence both downlink and uplink channels fade together. Closed loop power control still provides a gain in such a case but obviously the commands can be sent at a much slower rate than the FDD case. Figure 3

shows such a gain for a power control command rate of 100 and 200 commands per second. [5]

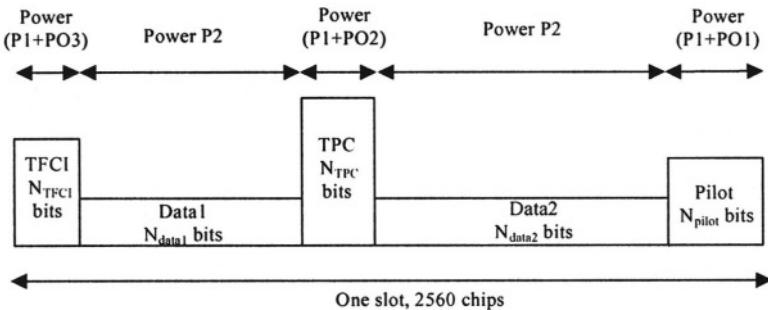


Figure 4: Down link slot format in the 3GPP standard (each frame is 15 slots)

6.4 Errors in the power control commands

Data bits are transmitted in frames. A frame is few msec long (10 msec or 20 msec). Data bits are coded and interleaved over one frame to give them more error protection. However, each frame has many slots (power control groups) where each of them has a unique transmit power control command (TPC command). These commands are sent uncoded to avoid the delay in the coding/decoding process, which increases their bit error rate (BER). To understand such issues more, Figure 4 shows the slot structure in the downlink of the 3GPP standard (UMTS). Few control fields are time multiplexed with the data bits. One of these fields is the transmit power control field (TPC) which can have N bits. The other filed that will have an impact in the power control algorithm is the TFCI field: transport format combination indicator. Data can be transmitted at different rates and the TFCI bits indicate at what rate the data is sent. The TFCI word is interleaved over the whole frame. One thing to notice here is the ability to send the down link control bits with extra power (offset from the data bits power) to reduce their error rate.

Two possible ways to decrease the TPC commands bit error rate (BER) are to increase the power of the TPC field or to increase the number of TPC bits to be transmitted in downlink. Figure 5 shows the BER probability for the TPC and the data bits (DPDCH) for different number of TPC bits and different power offsets. It is quiet clear from the figure how poor the TPC bits performance even when two bits are sent as a command (10%). The performance enhances when these commands are sent with extra powers. [6]

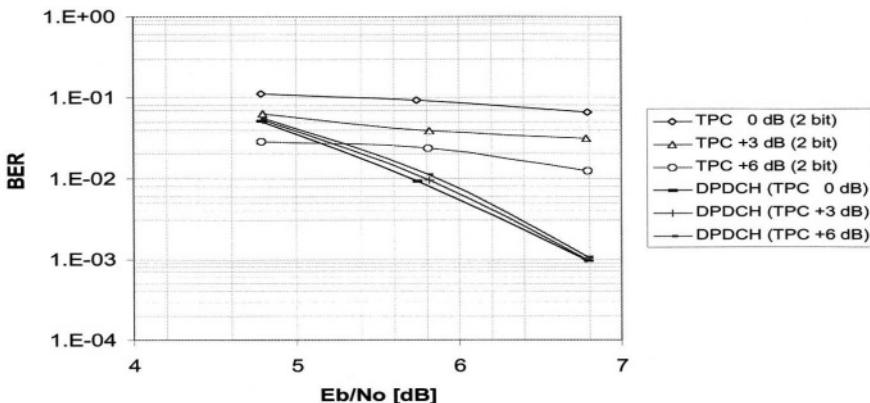


Figure 5: Downlink. TPC bit error rate as a function of the power offset in the TPC field

6.5 Power Control Step Size

One of the frequently asked questions is what is the optimum TPC step size? To answer such a question, we can look at Figure 6 which shows the required Eb/No for different step sizes. Of course, the lower the Eb/No , the better the step size. Conceptually, the small size should be small when the variations in the communications channel are small and vice versa. Looking at the figure, we see that the optimum step size is 0.5 dB when the channel is varying slowly (3 Km/h velocity). The optimum step size changes to 1 dB when the velocity changes to 10 Km/h. At a high velocity (100 Km/h), the step size is again 0.25 rather than being a big step size. This simply says that power control does not work at high speeds! As we have mentioned earlier, the power control does not adjust the power instantaneously but rather there is a delay in the loop. The assumed delay for the results in Figure 6 is one slot or one power control group. At high speeds, power control only creates more deviation in the received signal power. Hence, the larger the step size, the worse the situation becomes at high speeds. In other words, if the algorithm is not able to track the channel variations, it is better to turn it off.

Many papers are published investigating the gain from a variable step size and some suggest using also a delta modulation type of step size where a consecutive up (down) command results in larger step size. It might be hard to implement small step sizes in real life and hence these small step

sizes can be emulated. When the step size is 0.25, for example, the UE can wait for two consecutive down steps before reducing its power by 0.5 dB. Thus even though the UE reduced its power by 0.5 dB over a two slots period, it is similar to reducing the power by 0.25 in each slot.

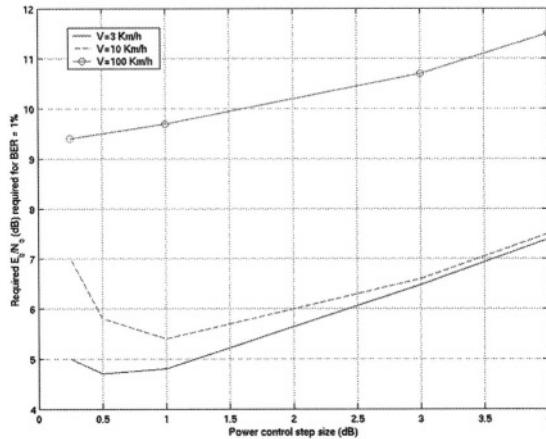


Figure 6: Received Eb/No performance of different step sizes at different velocities

6.6 Estimating Eb/No

Estimating the achieved Eb/No is one of the most tricky parts in the power control algorithm even though it looks like a straight down task. The problem arises from the fact that 3G systems supports multiple rate services and to measure Eb , the transmission rate has to be known. As we have mentioned earlier, in the 3GPP standard, each frame can have a different transmission rate and the TFCI word indicates this rate. However, to get the TFCI word, the 15 slots have to be received. That is to say, the whole frame has to be received before we can know the transmission rate of the frame. During the frame, however, 15 TPC commands have to be issued. Hence, waiting till the end of the frame is not an option. Fortunately enough, the control bits are sent at a known rate and hence they can be used to estimate the Eb/No . The number of the control bits is much less than the number of the data bits and using them to estimate the achieved Eb/No increases the measurement error. Other schemes to be used include using the dedicated pilot channel. The pilot channel is, however, not power controlled which creates another type of problems.

6.7 Outer loop power control

The typical target for different services is the block error rate (BLER). Unfortunately, there is no one-to-one mapping between the BLER and the required Eb/No . The required Eb/No is a function of many factors that include channel type, mobile velocity, etc. Hence the target Eb/No is controlled by the outer loop. The idea behind the outer loop power control algorithm is very simple. If the receiver (UE in case of down link power control and BTS in case of uplink power control) is receiving error free frames, the target should be reduced and if it receives frames in error, the target should be increased. Assuming that the current achieved Eb/No equals the target Eb/No . Then decreasing the target will create an error in the next frame. That is to say, the system will oscillate between the error free and erroneous case. This does not achieve the target BLER. To achieve the target BLER the up step and the down step for adjusting the Eb/No should be governed by the following relation

$$\text{Down_Step} = \text{Up_Step}/(1/\text{BLER} - 1) \quad (1)$$

That is to say, the Eb/No target is increased by a large step once a frame is found to be in error while the target is reduced slowly in case of error free transmission [8]. Of course, the above algorithm assumes that there are frames to be checked for erroneous! In case of DTX transmission (discontinuous transmission), the outer loop will not be able to do such checks and other means to estimate the link quality. One of such schemes is to use the bit error rate (BER) of the control bits. However, there is no one-to-one mapping between the BER of the control bits and the BLER of the data bits. Another issue here is if the UE should be allowed to control the target Eb/No for the down link power control? If the UE sets the target at the maximum allowed level, then the UE will be asking the BTS to send more and more power. This may make the link between the BTS and this particular UE error free but of course at the expense of excessive interference to the system.

6.8 Uplink Power control during soft handoff (SHO)

During SHO each BTS issues independently a TPC command to the UE. The UE can be receiving conflicting commands to adjust its power from the different BTSs. The “or of the downs” rule was used in the IS-95 standard to adjust the power in such a case. The rule is based on the idea of reducing the interference. If any base station is receiving extra power and asking the UE to decrease its power, the UE should decrease its power to reduce the

interference. This is the optimum combining rule assuming that all the received TPC commands are error free. When the mobile is in soft handoff with N base stations, it will be receiving N signals, which are combined optimally. Hence, some of the down link connections (base station to UE) can be strong while others can be weak. The TPC commands are, however, not combined. This results in considerable increase in the TPC BER in the case of SHO. So, in the “or of the downs” combining method, all the base stations can be asking for an increase in power and due to an error in only one of the power commands bits, the mobile decreases its power. To overcome such a problem, the UE has to do a reliability check on the TPC commands before combining them and only combine reliable ones. This reliability check can be achieved by measuring the signal to interference ratio of the link. Another way would be to assign a weight to each TPC command before combining them or to have a soft decision on the TPC command rather than a hard one. To reduce the interference, if any reliable link asks for a decrease in the transmitted power, the power is reduced. Another way to reduce the variation in the UE transmitted power during SHO is to reduce the power control step size. Figure 7 shows a comparison between the different combining schemes. The target Eb/No in these simulations was 5 dB. Using the “or of the downs” rule we see that the achieved Eb/No is greater than 12 dB is 90%. This is a strange result since we said that the UE will tend to reduce its power when the “or of the downs” rule is applied. To understand this we have to think of the outer loop function. In SHO, the BTSs start sending TPC commands to the UE which starts to reduce its power erroneously resulting in the frames being received in error. This triggers the outer loop that starts increasing the desired Eb/No target to high values resulting in the UE transmitting high power. That is to say, to avoid the impact of the erroneous behavior of the UE, the outer loop sets the Eb/No to high values. This is of course not helpful to the system capacity. We see from the figure that reducing the TPC step size to 0.25 dB (from 0.5 dB) still does not solve the problem. Ignoring the unreliable commands in the combining process clearly elevates the impact of the problem resulting in a performance close to the TPC error free case. [10]

6.9 Downlink power control during SHO

When the UE is in soft handoff (SHO), it combines the received signal from the different base stations that are in the active set and issues a single power command that all the base stations have to follow. The active set is the group of BTSs that are communicating with the UE simultaneously. The base stations in the active set demodulate the power command and adjust their powers independently. The mobile sends the power control command

on the uplink. The uplinks between the mobile and the different base stations are independent. This results in different instantaneous power control error rates on the different reverse links. Thus even though the mobile sends a single power control command, the base stations in the active set can adjust their powers in opposite directions.

The deviation in the base stations transmitted powers results in a loss of the diversity gain that we would otherwise have from SHO. Also, due to the errors in the power control commands, one of the base stations can unnecessarily transmit at a high power level that increases the interference. It

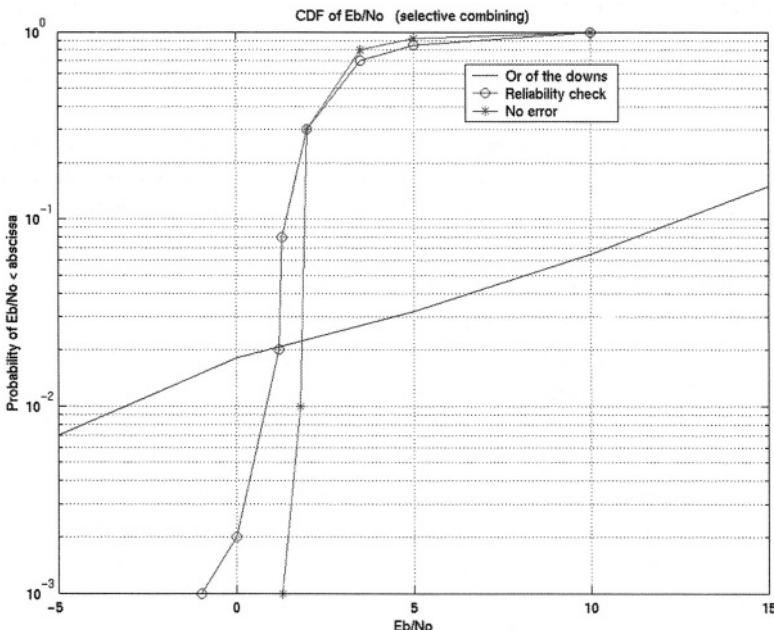


Figure 7: Downlink TPC commands combining during SHO (Desired Eb/No is 5dB)

could also erroneously transmit at a very low power and hence not contribute in the quality of the received signal.

To achieve the diversity gain, when in SHO, it is desirable to keep the base stations transmitted powers at comparable levels (not necessarily equal levels). One technique is to synchronize the base stations transmitted powers every few frames in order to make sure that their transmitted powers are as desired. This scheme however can be slow and requires signaling between the mobile and the base stations and a central controller. Another scheme for reducing the deviation in the base stations transmitted powers is to reduce the errors in the power control commands sent by the UE. By reducing the deviation, the requirement on synchronization frequency can be loosened.

The reduction in transmitted powers can be achieved by giving these commands extra power (power offset from the data bits). However, this scheme might not be a feasible option due to the power amplifier non-linearity problems at the UE. As an another option, the TPC commands error rate can be reduced by repeating the same power control command over more than one power control group.

6.9.1 Base Stations Transmitted Powers Synchronization

The base stations in the active set have to report their transmitted power levels to a central controller every M frames and the controller has to resynchronize their power levels. The central controller will send a new power level at which all the base stations in the active set have to transmit. This enforces all the base stations in the active set to transmit at equal power levels every M frames. There are different ways for the central controller to determine the new power level. The new power level can simply be the average of the reported base stations powers. A better algorithm will try to determine which reported power level is the closest to what the UE actually wants. This can be done by observing the quality of the uplink. The UE sends the TPC commands over the uplink and hence a better uplink results in fewer errors on the TPC commands. The fewer the errors on the TPC commands, the closer the transmitted power to what the UE requires. The reported uplink quality can be the BLER or the signal to interference ratio (SIR) or any other measurements that can indicate which link has lower TPC error rate.[11]

6.9.2 TPC rate reduction

The TPC rate reduction is applied to reduce the TPC error rate and hence the deviation in the BTSs transmitted powers is reduced. When in soft handover, the BTSs keep adjusting their power levels every slot. Assuming the frame to have 15 slots, this results in the BTSs adjusting their powers 15 times/frame. If the step size is 1dB and two BTSs adjust their powers in an opposite direction 4 times, this results in an 8 dB difference in their transmitted powers. This obviously will waste the diversity gain that we achieve by having the UE in soft handoff. When the UE is in soft handoff, it is receiving the signal from more than one BTS and probably via more than one fading path from each BTS. Thus, the high rate of 1500 commands may not be needed when in soft handoff. Reducing the power control rate during

soft handoff results in the BTSs adjusting their transmitted powers less often which reduces the deviation in their transmitted powers.

Another advantage of reducing the power adjustment rate is giving more error protection to the TPC commands sent by the UE. Currently, the UE sends one power control command every slot. Another option would be that the UE sends one power control command every K slots. This means that the BTSs in soft handoff with the UE adjusts their transmitted powers only every K slots. The simplest way will be to repeat the same command over the K slots. However, this is obviously not the optimum way if the command is a multi step command in terms of error protection. Simple coding can be employed in such a case depending on the power control adjustment rate. This also provides some interleaving which will lower the power control command error rate. To achieve this the 3GPP standard supports two modes for uplink power control operation. The UE checks the downlink power control mode (DPC_MODE) before generating the TPC command: [2]

if DPC_MODE = 0 : the UE sends a unique TPC command in each slot and the TPC command generated is transmitted in the first available TPC field in the uplink DPCCH;

if DPC_MODE = 1 : the UE repeats the same TPC command over 3 slots and the new TPC command is transmitted such that there is a new command at the beginning of the frame.

Upon receiving the TPC commands, the BTS shall adjust its downlink power accordingly. For DPC_MODE = 0, BTS shall estimate the transmitted TPC command to be 0 or 1, and shall update the power every slot. If DPC_MODE = 1, BTS shall estimate the transmitted TPC command over three slots to be 0 or 1, and shall update the power every three slots.

After estimating the k:th TPC command, BTS shall adjust the current downlink power $P(k-1)$ [dB] to a new power $P(k)$ [dB] according to the following formula:

$$P(k) = P(k - 1) + P_{TPC}(k) + P_{bal}(k), \quad (2)$$

where $P_{TPC}(k)$ is the k th power adjustment due to the inner loop power control, and $P_{bal}(k)$ [dB] is a correction according to the downlink power control procedure for balancing radio link powers towards a common reference power. The power balancing procedure is described later.

Simulations are shown in Figure 8 that show the impact of reducing the rate of the TPC commands from 1500 commands per second to 500

commands per second. The rate reduction is achieved by repeating the same TPC command over three slots. [11] It can be seen that the rate reduction algorithm reduces the probability that the deviation in the transmitted powers exceeding 5 dB from 20% to 2%. This is obviously a big gain which can ensure that UEs in SHO are having the desired diversity gain. As was the case in the downlink, a small TPC step size is better for the uplink power control also when in SHO. The impact of the step size (δ) is shown in Figure 9 [14]

6.9.3 Adjustment loop

Adjustment loop is another scheme that is used to reduce the deviation in the BTSSs transmitted powers during SHO. For adjustment loop, DL reference power P_{REF} and DL power convergence coefficient r ($0 < r < 1$) are set in the active set BTSSs during soft handover so that the two parameters are common to the BTSSs. For simplicity, DL powers of two BTSSs are considered in this explanation.

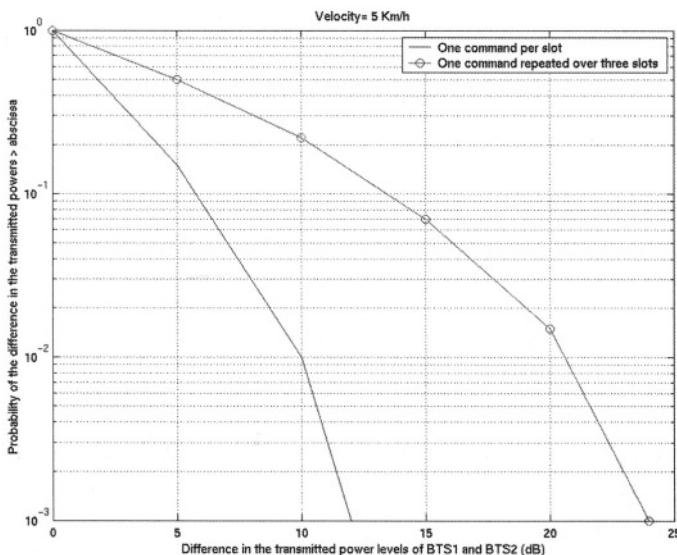


Figure 8: Deviation in the BTSSs transmitted powers

Adjustment loop works in addition to inner loop power control, and DL power at slot i of two BTSSs, $P1(i)$, and $P2(i)$, are updated at a certain interval (typically in every slot as in this explanation) as follows:

$$P1(i+1) = P1(i) + (1 - r)(P_{REF} - P1(i)) + S_{INNERLOOP1}(i) \quad (3)$$

$$P2(i+1) = P2(i) + (1 - r)(P_{REF} - P2(i)) + S_{INNERLOOP2}(i) \quad (4)$$

Where $S_{INNERLOOP1}$ and $S_{INNERLOOP2}$ are the inner loop power adjustments according to the received TPC commands. The difference in the BTSSs transmitted powers is derived from equations (1) and (2) if TPC error does not occur i.e. $SINNERLOOP1(i)$ and $SINNERLOOP2(i)$ are equal.

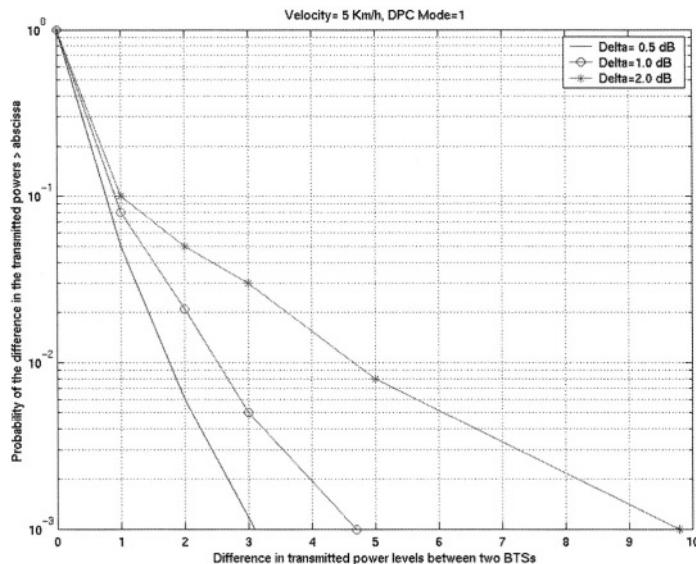


Figure 9: Impact of the step size (0.5 dB, 1dB and 2 dB) on the deviation between two BTSSs transmitted powers

$$P1(i+1) - P2(i+1) = r(P1(i) - P2(i)) = r^i(P1(1) - P2(1)) \quad (5)$$

Equation (5) means that the difference converges at zero when r is smaller than one. The gain from this algorithm is limited by the fact that the BTS power can not be adjusted by tiny steps but still it is shown that such scheme can help in reducing the deviation in the BTSSs transmission power [16].

6.10 Site Selection Diversity TPC (SSDT)

SSDT is an extreme case of downlink power control in a SHO region where the UE selects only one BTS to transmit at a time. SSDT operation is summarized as follows. The UE selects one of the BTSSs from its active set to

be ‘primary’, all other BTSs are classed as ‘non primary’. The main objective is to transmit on the downlink from the primary BTS, thus reducing the interference caused by multiple transmissions in a soft handoff mode. A second objective is to achieve fast site selection without network intervention, thus maintaining the advantage of the soft handoff. In order to select a primary BTS, each BTS is assigned a temporary identification (ID) and UE periodically informs a primary BTS ID to the connecting BTSs. The non-primary BTSs selected by UE switch off the transmission power. The primary BTS ID is delivered by UE to the active BTSs via uplink feedback bits. SSDT activation, SSDT termination and ID assignment are all carried out by higher layer signaling. UE may also be commanded to use SSDT signaling in the uplink although BTSs would transmit the downlink without SSDT being active. In case SSDT is used in the uplink direction only, the processing in the UE for the radio links received in the downlink is as with macro diversity in non-SSDT case. Higher layers set the downlink operation mode for SSDT. [2]. Simulations in [17] show the gain from SSDT over conventional TPC. A sample of these results is shown in Table 1 to get a feeling of how large this gain could be. The gain from SSDT is achieved at low speeds where the UE is capable of exploiting the multipath fading diversity between multiple cells.

Table 1: Capacity [kb/s/MHz/Sector] of SSDT and conventional TPC for pedestrian UE in terms of number of RAKE fingers

| | Infinite RAKE fingers | 6 RAKE fingers | 5 RAKE fingers |
|------------------|-----------------------|----------------|----------------|
| Conventional TPC | 377 | 233 | 217 |
| SSDT | 479 | 361 | 341 |

6.11 Reduced active set

As we have seen in the previous section, site selection might be an alternative to the typical implementation of soft handoff in CDMA systems. One question here is if the data should be available at all the BTSs in the active set and be transmitted immediately when the UE selects the primary BTS or not. If the answer is yes, this will require more buffering at the BTSs. If the answer is no, the data should be sent from a central controller to the selected (primary) BTS. This of course causes more delay. This is one of the reasons why voice and data tolerating delay can be transmitted from a different group of cells. Voice clearly requires SHO more than data due to its delay intolerance. Data applications usually employ retransmission schemes

(ARQ) that allows receiving the data without errors even if the first transmission is erroneous. This lead to the concept of reduced active set. The concept simply says that there can be the regular active set where all the BTSs in the set transmits one service (voice for example) and a reduced number of these BTSs can transmit another service to the same UE. In the above case the communication channel will be different for the two services. This is because the two channels will be arriving through different paths (different BTSs). The power control algorithm has been modified in the CDMA2000 system to allow controlling the power of both channels independently. In IS200 the voice channel is called the fundamental channel while the data channels are referred to as supplemental channels. The power control bits are time multiplexed to allow controlling both fundamental and supplemental channels differently. In IS200, the TPC rate is 800 commands/sec. These commands can be split between the fundamental channel and the supplemental channel in three different ways: 800/0, 400/400 and 600/200.

6.12 Compressed mode

Third generation systems are expected to replace second generation systems in a gradual way. Operators may deploy 3G systems in urban areas first leaving the 2G systems covering the suburb areas. Hence, a UE should be able to make a hard handoff from 3G systems to 2G systems. To make such a handoff, the UE should be able to measure the signal strength of the 2G carriers. This is an example where inter-frequency measurements is required (a measurement on a carrier operating on a different frequency than the current carrier frequency).

There are two solutions for such an issue. First, the UE should have two receivers allowing it to do measurements on a separate carrier while receiving data on the current carrier. This is not a highly desirable solution due to the UE complexity increase. The other solution is to create a gap in the transmission allowing the UE to do the required measurements during this gap. This is referred to as the compressed mode or the slotted mode in the 3GPP standard. The interruption of the transmission creates a break in the operation of the power control algorithm. The inner loop requires receiving data to be able to estimate the achieved Eb/No and needs to issue TPC continuously to track the communication channel variations. Figure 10 shows how much a Raleigh fading channel can change during a gap of 8 slots. For a Doppler frequency of 40 Hz, the probability that the change being more than 5 dB is about 10%. Detailed procedures on how the power control algorithm should work after resuming the transmission (after the

gap) can be found in [2]. The main principle in these procedures is to use a larger step size to adjust the Eb/No target faster after the gap to adjust the power in a faster way compared to the normal mode. Looking at these procedures, it is clear how important continuous transmission is for the operation of the power control algorithm. Power control algorithms and especially the closed loop were not designed for discontinuous transmission.

6.13 Power/Rate control

Even though the control of the rate in addition to the power is not a feature of early releases of 3G standards, it should be mentioned here since it is included in enhanced releases and it has large impact of the way we think of power control algorithms. Enhanced releases of 3G systems employ adaptive modulation/coding (ACM) for data transmission. As we have mentioned very early in this chapter, the objective of the TPC algorithm is to make sure that the received Eb/No is equal to the desired Eb/No , not less or more. Let us define Eb/No .

$$Eb/No = (S/R)/[(I+N)/W] \quad (6)$$

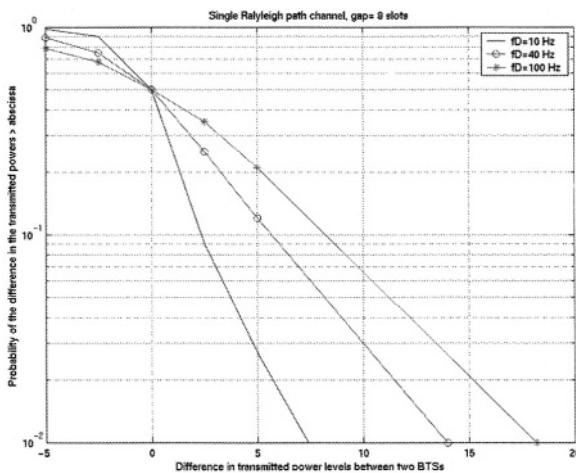


Figure 10: Complementary CDF of the change in the communication channel over 8 slots, the channel is a single path Rayleigh fading channel (fD is the Doppler frequency)

Where S is the signal power, R is the transmission rate, I is the interference level, N is the white noise and W is the system bandwidth. In typical power control algorithms, the power of the signal (S) is adjusted. This is obviously not the only parameter that can be adjusted to achieve the

required Eb/No . The transmission rate (R) can be also adjusted where a lower transmission rate results in a higher Eb/No . This of course assumes that the service can tolerate the delay associated with reducing the transmission rate. Considering a UE in a deep fade. Typical power control algorithms increases the UE transmission powers which penalize all the other users by causing more interference. However, reducing the UE transmission rate penalize the affected UE by reducing its transmission rate.

Rather than changing the rate directly, ACM can be employed to change the required Eb/No . High speed data packet access channel (HSDPA) in 3GPP is an example where the channel is transmitted at a constant power. Served UEs measure the signal to interference ratio (SIR) of the downlink and report it back to the BTS. The BTS selects an ACM that allows the UE to receive the data at a desired BLER. The ACM levels include different modulation levels (QPSK, 16 QAM, 64 QAM, etc) and different coding rates (3/4, 1/2, 1/3, etc) This of course means that UEs closer to the BTS will have a higher data rate compared to users at the edge of the cell. To increase the system throughput, a scheduler is employed at the BTS that serves the users who have a good SIR and provides fairness between the different users by providing longer time for the disadvantaged users.

It might look like that the ACM concept is completely different from the power control algorithm. However, there are lots of similarities between the two schemes; after all both algorithms try to achieve the Eb/No that satisfies a certain BLER. Take as an example the outer loop. There is still no one to one mapping between the reported SIR and the required ACM level to achieve a certain BLER. An outer loop is still required for the ACM operation. Hence, many of the lessons learned from power control algorithms can be extended to ACM design.[21]-[23].

REFERENCES

1. Hashem B. and Yanikomeroglu H. "Power Control for Third Generation Cellular Systems ". The International Congress on Dynamics and Control of Systems, DYCONS99, Ottawa, Canada
2. 3GPP standard documents: TS 25.214 V4.1.0, "Physical layer procedures (FDD)"
3. Hashem B. and Sousa E., "Reverse Link Capacity and Interference Statistics of Fixed Step Power Controlled DS/CDMA System Under Slow Multipart Fading", IEEE Transactions on Communications, December 1999

4. Jiangzhou Wang, "Open-loop power control error in cellular CDMA overlay system", IEEE Journal on Selected Areas in Communications, Volume: 19 Issue: 7 , July 2001 Page(s): 1246–1254
5. 3GPP standard documents: TSGR1(00)0934, " The performance improvement from power control"\
6. 3GPP standard documents: TSGW1 #3(99)171, "BER on power control bits"
7. 3GPP standard documents: TSGR1#6(99)821, "Optimum Power Control Step Size in Normal Mode"
8. Sampath, A.; Sarath Kumar, P.; Holtzman, J.M. On setting reverse link target SIR in a CDMA system Vehicular Technology Conference, 1997, IEEE 47th , Volume: 2 , 1997 Page(s): 929 -933 vol.2
9. 3GPP standard documents: TSGR1 #3(99)213, "Potential downlink power control instabilities during soft handover"
10. Hashem B. and Secord N., "Combining of Power Commands Received from Different Base Stations during Soft Handoff in Cellular CDMA Systems", International Symposium on Personal, Indoor and Mobile Radio Communications PIMRC '99 Osaka, Japan
11. Hashein B. and Le Strat E. "On the Balancing of the Base Stations Transmitted Powers During Soft Handoff in Cellular CDMA Systems", ICCC2000, New Orleans, Louisiana, USA June 18-22, 2000
12. 3GPP standard documents: TSGR1#7 99bl5, "Downlink Power Control Rate Reduction during Soft Handover"
13. Furukawa, H.; Harnage, K.; Ushirokawa, A. " SSDT-site selection diversity transmission power control for CDMA forward link" IEEE Journal on Selected Areas in Communications, Volume: 18 Issue: 8 , Aug. 2000 , Page(s): 1546–1554
14. 3GPP standard documents: TSGR1 #9(99)k02, "Downlink Power Control Step Size During Soft Handover"
15. 3GPP standard documents: TSGR1#7bis(99)E69, "Adjustment Loop in downlink power control during soft handover"
16. Hamabe, K , Adjustment loop transmit power control during soft handover in CDMA cellular systems, Vehicular Technology Conference, 2000. IEEE VTS Fall VTC 2000. 52nd , Volume: 4 , 2000 Page(s): 1519-1523 vol.4
17. Takano, N.; Hamabe, K, Enhancement of site selection diversity transmit power control in CDMA cellular systems, Fall. IEEE VTS 54th Vehicular Technology Conference, 2001. VTC 2001, Volume: 2, 2001, Page(s): 635 –6393
18. GPP standard documents: TSGR1 #4(99)459, "Effect of slotted mode on the power control algorithm"

- 19.3GPP standard documents: TR25.848 (Technical Report on Physical Layer Aspects of UTRA High Speed Downlink Packet Access Channel)
20. Lei Song; Mandayam, N.B.; Gajic, Z, Analysis of an up/down power control algorithm for the CDMA reverse link under fading , IEEE Journal on Selected Areas in Communications, Volume: 19 Issue: 2 , Feb 2001, Page(s):277 –286
- 21.Chi Wan Sung; Wing Shing Wong, Power control and rate management for wireless multimedia CDMA systems, IEEE Transactions on Communications, Volume: 49 Issue: 7 , July 2001 , Page(s): 1215–1226
- 22.Mandyam, G.D.; Yi-Chyun Tseng, Packet scheduling in CDMA systems based on power control feedback Communications, IEEE International Conference on 2001. ICC 2001. Volume: 9 , 2001 Page(s): 2877 -2881 vol.9
- 23.Bender, P.; Black, P.; Grob, M.; Padovani, R.; Sindhushyana, N.; Viterbi, S. CDMA/HDR: a bandwidth efficient high speed wireless data service for nomadic users IEEE Communications Magazine, Volume: 38 Issue: 7 , July 2000, Page(s): 70 -77

This page intentionally left blank

Chapter 7

POWER CONTROL IN WIRELESS NETWORKS

Characteristics and Fundamentals

Fredrik Gunnarsson

Dept. of Electrical Engineering

Linköpings universitet

SE-581 83 Linköping, Sweden

fred@isy.liu.se

Abstract The global communications systems critically rely on control algorithms of various kinds. In UMTS (universal mobile telephony system) – the third generation mobile telephony system just being launched, power control algorithms play an important role for efficient resource utilization. This chapter describes power control fundamentals including both theoretical and practical limitations. The relations to session management such as admission and congestion control is also addressed. Concepts and algorithms are illustrated by simple examples and simulations.

Keywords: power control, wireless networks, distributed control, stability, time delays, limitations, disturbance rejection, measurement errors, estimation errors.

Introduction

While the demand for access to services in wireless communications systems is exponentially growing, an increased interest in utilizing the available resources efficiently can be observed. A consequence of the limited availability of radio resources is that the users have to share these resources. Power control is seen as an important means to reduce mutual interference between the users, while compensating for time-varying propagation conditions. The objective with this chapter is to introduce power control algorithms and to point at some central and critical issues partly using a control theory framework.

Power control has been an area subject to extensive research in recent years. Some surveys include (Rosberg and Zander, 1998; Hanly and Tse, 1999; Gunnarsson and Gustafsson, 2002). A simplified radio link model is typically

adopted to emphasize the network dynamics of power control. The transmitter is using the power $\bar{p}(t)$ to transmit data scrambled by a user-specific code. The channel is characterized by the power gain $\bar{g}(t) (< 1)$, where the “bar” notation indicates linear scale, while $g(t) = 10 \log_{10}(\bar{g}(t))$ is in logarithmic scale (dB). Correlating the received signal with the code, the receiver extracts the desired signal, which has the power $\bar{C}(t) = \bar{p}(t)\bar{g}(t)$, and is also subject to an interfering power $\bar{I}(t)$ from other connections. The perceived quality is related to the *signal-to-interference ratio (SIR)* $\bar{\gamma}(t) = \bar{p}(t)\bar{g}(t)/\bar{I}(t)$. For error-free transmission (and if the interference can be assumed Gaussian), the achievable data rate $R(t)$ is limited from above by (Shannon, 1956)

$$R(t) = W \log_2 (1 + \bar{\gamma}(t)) \text{ [bits/s]}, \quad (7.1)$$

where W is the bandwidth in Hertz. From a link perspective, the objective with power control can for example be

- to use constant power and variable coding and modulation to benefit from times of a favorable channel
- to maintain constant SIR and thereby constant data rate in lieu of (7.1)

The former is information theoretically sound, and is further explored in e.g. (Goldsmith, 1997). In this chapter, the focus is on networks with relatively slow rate adaption mechanisms. Therefore the aim is the latter objective with power control to mitigate power gain and interference variations to maintain constant SIR and thereby constant rate. Figure 7.1 illustrates the two fundamentally different objectives.

Necessary system models and the notation are introduced in Section 1. Power control is also described, starting from an integrating controller, which is related to important proposals in the literature. Section 2 provides fundamental limitations of this controller on link level. Further network limitations are addressed in the subsequent section together with power control convergence results. Since power control is instrumental to other related resource management algorithms, a smorgasboard of different aspects of power control is brought up in Section 4.

1. System Model

1.1 Power Gain

By neglecting data symbol level effects, the communication channel can be seen as a time varying power gain made up of three components $g(t) = g_p(t) + g_s(t) + g_m(t)$ as illustrated by Figure 7.2. The signal power decreases with distance \bar{d} to the transmitter, and the path loss is modeled as $g_p = K_p - a_p \log_{10}(\bar{d})$ (Okamura et al., 1968). Terrain variations cause diffraction phenomena and this shadow fading g_s is modeled as ARM $A(n_a, n_b)$ -filtered Gaussian white noise (n_a is typically 1-2, $n_b = n_a - 1$, (Sørensen,

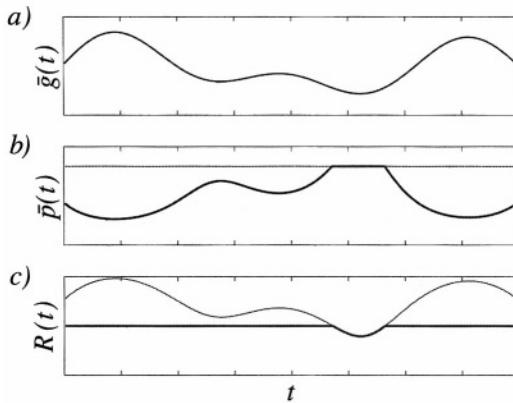


Figure 7.1. Illustration of power control objectives when subject to a varying power gain (a). b) Using a varying power (thick line) to mitigate channel variations or using maximum power (thin line) result in c) a constant data rate as long as less than maximum power is required (thick line) and a varying data rate depending on the channel (thin line) respectively.

1998)). The multipath model considers scattering of radio waves, yielding a rapidly varying gain g_m (Sklar, 1997).

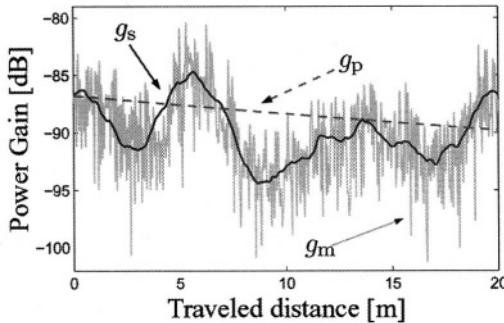


Figure 7.2. The power gain $g(t)$ is modeled as the sum of three components: path loss $g_p(t)$, shadow fading $g_s(t)$ and multipath fading $g_m(t)$. Here this is illustrated when moving from a reference point and away from the transmitter.

A 3GPP Typical Urban (3GPP, 2003a) fast fading channel model will be utilized throughout the chapter. It models a multipath channel and a receiver unable to utilize all paths. Power gain values over 20 m are depicted in Fig. 7.3a. The frequency content of the power gain can be described independent of mobile velocity by expressing it with respect to the spatial frequency $\xi [\text{m}^{-1}]$. As seen in Fig. 7.3b, most of the frequency content is concentrated below 60 m^{-1} . For example the velocity $v \text{ m/s}$ means that the disturbance energy is concentrated below $60v \text{ Hz}$. Note that this is much higher than the Doppler frequency of

each individual path (Jakes, 1974). This is due to the combination of several paths in the receiver.

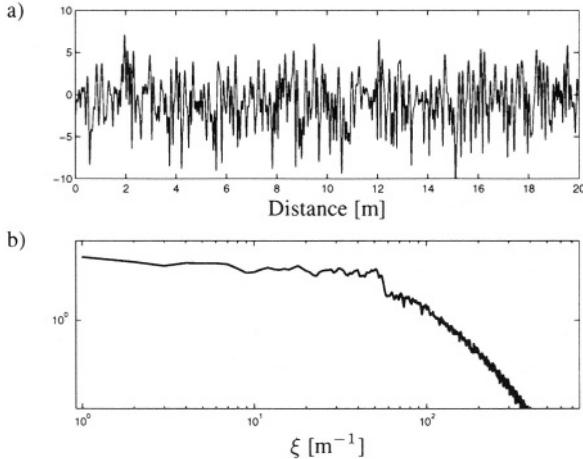


Figure 7.3. Realization of 3GPP Typical Urban power gain in dB a) with respect to traveled distance and b) in the frequency domain w.r.t. spatial frequency.

1.2 Wireless Networks

Consider a general network with m transmitters using the powers $\bar{p}_i(t)$ and m connected receivers. For generality, the base stations are seen as multiple transmitters (downlink) and multiple receivers (uplink). The signal between transmitter j and receiver i is attenuated by the power gain \bar{g}_{ij} . Thus receiver i connected to transmitter i will experience a desired signal power $\bar{C}_i(t) = \bar{p}_i(t)\bar{g}_{ii}(t)$ and an interference from other connections plus noise $\bar{I}_i(t)$. The *signal-to-interference ratio (SIR)*¹ at receiver i can be defined by

$$\bar{\gamma}_i(t) = \frac{\bar{C}_i(t)}{\bar{I}_i(t)} = \frac{\bar{g}_{ii}(t)\bar{p}_i(t)}{\sum_{j \neq i} \bar{g}_{ij}(t)\bar{p}_j(t) + \bar{\nu}_i(t)}, \quad (7.2)$$

where $\bar{\nu}_i(t)$ is thermal noise at receiver i .

Depending on the receiver design, propagation conditions and the distance to the transmitter, the receiver is differently successful in utilizing the available desired signal power $\bar{p}_i\bar{g}_{ii}$. Assume that receiver i can utilize the fraction $\bar{\delta}_i(t)$ of the desired signal power. Then the remainder $(1 - \bar{\delta}_i(t))\bar{p}_i\bar{g}_{ii}$ acts as interference, denoted *auto-interference* (Godlewski and Nuaymi, 1999). We will assume that the receiver efficiency changes slowly, and therefore can be considered constant. Hence, the SIR expression in Equation (7.2) transforms

to

$$\bar{\gamma}_i(t) = \frac{\bar{\delta}_i \bar{g}_{ii}(t) \bar{p}_i(t)}{\sum_{j \neq i} \bar{g}_{ij}(t) \bar{p}_j(t) + (1 - \bar{\delta}_i) \bar{p}_i(t) \bar{g}_{ii}(t) + \bar{\nu}_i(t)}. \quad (7.3)$$

From now on, this quantity will be referred to as SIR. For efficient receivers, $\bar{\delta}_i = 1$, and the expressions (7.2) and (7.3) are equal. In logarithmic scale, the SIR expression becomes

$$\gamma_i(t) = p_i(t) + \delta_i + g_{ii}(t) - I_i(t). \quad (7.4)$$

It is also customary to use an alternative notation, where index i identifies the mobile, index j a base station, and j_i the base station mobile i is connected to. In addition, the distinction between uplink and downlink can be made more clear.

In the downlink, many of the interfering signals go through the same channels. It can therefore be motivated to denote the power gain from base station j to mobile i by \bar{g}_{ij}^{DL} . Let $\bar{P}_j(t)$, $j = 1, \dots, B$ denote the total powers of the B base stations. Furthermore, the interfering signal powers from the same base station are attenuated by $\bar{\alpha}_i$ due to the effect of the channelization codes, which are orthogonal on the transmission side. Then, the alternative downlink SIR expression is given by

$$\bar{\gamma}_i^{\text{DL}}(t) = \frac{\bar{\delta}_i \bar{g}_{ij_i}^{\text{DL}}(t) \bar{p}_i(t)}{\bar{\alpha}_i(\bar{P}_{j_i}(t) - \bar{\delta}_i \bar{p}_i(t)) \bar{g}_{ij_i}^{\text{DL}}(t) + \sum_{j \neq j_i} \bar{P}_j(t) \bar{g}_{ij}^{\text{DL}}(t) + \bar{\nu}_i(t)}. \quad (7.5)$$

This expression can for example be used together with a downlink load assumption in terms of base station powers $\bar{P}_j(t) = \hat{\bar{P}}_j$ to analyze the downlink situation, perhaps conducted in a cell planning phase.

The uplink signals go through independent channels. However, it can still be instructive to use an alternative notation, where the power gain from mobile i to base station j is denoted \bar{g}_{ij}^{UL} . If the uplink and downlink channel were reciprocal, then $\bar{g}_{ij}^{\text{UL}} = \bar{g}_{ij}^{\text{DL}}$. This uplink power gain notation requires an alternative uplink SIR definition

$$\bar{\gamma}_i^{\text{UL}}(t) = \frac{\bar{\delta}_i \bar{g}_{ij_i}^{\text{UL}}(t) \bar{p}_i(t)}{\sum_{k \neq i} \bar{g}_{kj_i}^{\text{UL}}(t) \bar{p}_k(t) + (1 - \bar{\delta}_i) \bar{p}_i(t) \bar{g}_{ij_i}^{\text{UL}}(t) + \bar{\nu}_{j_i}(t)}. \quad (7.6)$$

Note that the power gain matrix $\bar{\mathbf{G}} = [\bar{g}_{ij}]$ is square $M \times M$, unlike the matrices $\bar{\mathbf{G}}_{\text{DL}} = [\bar{g}_{ij}^{\text{DL}}]$ and $\bar{\mathbf{G}}_{\text{UL}} = [\bar{g}_{ij}^{\text{UL}}]$, which are $M \times B$. The generic SIR expression (7.3) can be used through

$$\bar{g}_{ik} = \bar{g}_{ijk}^{\text{DL}}, \quad \bar{g}_{ik} = \bar{g}_{kji}^{\text{DL}}$$

respectively.

1.3 Fundamental Power Control Algorithm

We adopt the log-linear power control model in (Dietrich et al., 1996; Blom et al., 1998) and consider all quantities in logarithmic scale (dB). With the power control objective to maintain constant SIR irrespective of time-varying power gain and interference power, it is natural to estimate SIR $\gamma_i(t)$ in the receiver and compare to the SIR reference $\gamma_i^t(t)$. The following integrating controller adjusts the power as long as the control error $e_i(t) = \gamma_i^t(t) - \gamma_i(t)$ is non-zero

$$p_i(t+1) = p_i(t) + \beta (\gamma_i^t(t) - \gamma_i(t)) = p_i(t) + \beta e_i(t), \quad (7.7)$$

where the time index t refers to instants of power level updates in the transmitters. The linear control algorithm together with the SIR expression in (7.4) form a log-linear local control loop with the power gain, interference power and receiver inaccuracy as an additive disturbance. The local dynamical behavior can be conveniently described using a control theory framework. Introduce the *time-shift operator*² q as

$$q^{-n}p(t) = p(t-n), \quad q^n p(t) = p(t+n) \quad (7.8)$$

The integrating control algorithm in (7.7) can thereby be rewritten as

$$p_i(t) = \frac{\beta}{q-1} \hat{e}_i(t) = R(q) \hat{e}_i(t) \quad (7.9)$$

Hence, the power control local loop is described by the block diagram in Figure 7.4.

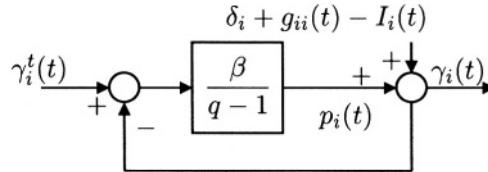


Figure 7.4. The local loop dynamics when employing the integrating controller in (7.9).

1.4 Relation to the Literature

The foundations for the distributed power control algorithms were laid in (Meyerhoff, 1974). The motivation in this article was not to obtain a distributed algorithm, but rather to make the centralized power control algorithms in Section 3 more computationally efficient. The proposed algorithms were based on results for iterative computations of eigenvectors.

These results are interpreted in (Grandhi et al., 1994; Foschini and Miljanic, 1993) as distributed power control utilizing only local information, and provided convergence results for the linear scale algorithm Distributed Power Control (DPC)

$$\bar{p}_i(t+1) = \bar{p}_i(t) \frac{\bar{\gamma}^t}{\bar{\gamma}_i(t)}. \quad (7.10)$$

The algorithms proposed in (Yates, 1995; Lee and Lin, 1996) aims at more careful control action compared to the DPC algorithm in (7.10). They can be rewritten as

$$\bar{p}_i(t+1) = \bar{p}_i(t) \left(\frac{\bar{\gamma}^t}{\bar{\gamma}_i(t)} \right)^\beta. \quad (7.11)$$

In logarithmic scale, this is (apart from individual target SIR's, $\bar{\gamma}_i^t(t)$) analogous to the integrating controller in (7.7), and DPC represents the specific case $\beta = 1$.

2. Fundamental Local Limitations

The power control loop in Figure 7.4 is essentially a block diagram describing the integrating controller together with theoretical and simplistic models. In practice, performance is fundamentally limited by several things as illustrated by Figure 7.5 (Gunnarsson, 2001).

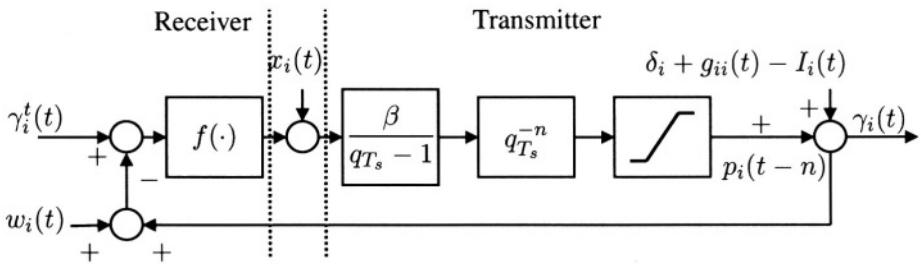


Figure 7.5. The local loop dynamics considering fundamental limitations and a realistic separation of the control loop into one receiver part and one transmitter part.

The limitations include that

- The **update rate** has a direct relation to the controller's ability of mitigating time-varying disturbances. It is indicated by the sample interval T_s together with the time shift operator q_{T_s} .
- The time-varying **measurement errors** $w_i(t)$ and **disturbances** $g_{ii}(t)$, $I_i(t)$ and δ_i make it difficult to equalize SIR.

- The **feedback bandwidth** limits the control information that is fed back from the receiver. It can be seen as quantization of the power control error $f(e_i(t))$.
- The **time delays** $q_{T_s}^{-n}$ effect the **local stability** of the control algorithm.
- The **transmitter power** is limited both from above and below.
- The feedback channel is subject to **feedback errors** $x_i(t)$, which further reduces the information about the current receiver situation.

If neglecting non-linear components due to quantization and saturation, the remaining linear loop can be analyzed using linear systems theory. Clearly, the power control objective is to maintain $\gamma_i(t) = \gamma_i^t(t)$ or equivalently $e_i(t) = 0$ despite limitations. From the block diagram in Fig. 7.5, we obtain

$$\gamma_i(t) = G(q)\gamma_i^t(t) + S(q)(g_{ii}(t) - I_i(t)) + G(q)w_i(t) \quad (7.12a)$$

$$e_i(t) = S(q)(\gamma_i^t(t) - g_{ii}(t) + I_i(t)) + G(q)w_i(t), \quad (7.12b)$$

where the dynamics is described by

$$G(q) = \frac{\beta}{q^n(q-1) + \beta}, \quad S(q) = \frac{q^n(q-1)}{q^n(q-1) + \beta} \quad (7.13)$$

$G(q)$ is referred to as the *closed-loop system*, and $S(q)$ as the *sensitivity function*. Note that $S(q) = 1 - G(q)$. The corresponding relations in the frequency domain are obtained by replacing q by $e^{i\omega T_s}$. The closed-loop system describes the tracking capability of the control algorithm, while the sensitivity function relates to the disturbance suppression performance. The effects of measurement (sensor) errors, however, is also captured by the closed-loop system. Furthermore, local loop stability is related to properties of $G(q)$.

2.1 Limited Update Rate

The actual time between consecutive power updates, the *sample interval* T_s , varies from systems to system. For example $T_s = 0.48$ s in GSM and $T_s = 1/1500$ s in WCDMA. To avoid confusion, we let the time index t represent instants of power level updates in the transmitters. Seemingly, this notation is equal to the assumption of synchronous updates, but the only needed assumption is that all transmitters update their power levels within the time frame of one sample interval.

Consider the ideal integrating controller in (7.7) with $\beta = 1$, update rate of 1500 Hz as in WCDMA and two mobiles with velocities 2.5 m/s and 12 m/s. Performance is evaluated in terms of control error $\gamma_i^t(t) - \gamma_i(t)$ over time. The power gain has a bandwidth of 60v Hz as described in Section 1.1, which

yields that the disturbance energy for the two mobiles is concentrated to up to 150 Hz and 720 Hz respectively. Both these frequencies are below the Nyquist frequency (750 Hz) and can thus be represented without alias. However, as seen in Figures 7.6a,b only the disturbance of the first mobile is compensated for. The answer lies in the sensitivity function in Figure 7.6c, where we note that only frequencies up to ≈ 200 Hz are suppressed.

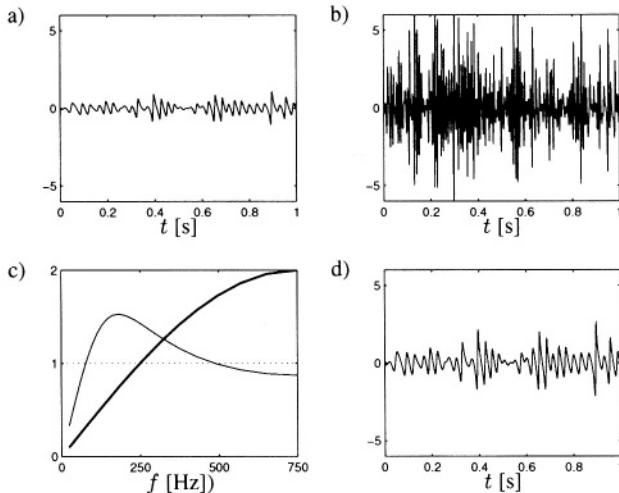


Figure 7.6. Performance of the integrating controller for $\beta = 1$ and no delay (a-c), and $\beta = 0.34$ and one slot delay (c-d). a) $e_i(t)$, slow mobile ($v=2$ m/s), b) $e_i(t)$, fast mobile ($v=12$ m/s), c) $|S(e^{i\omega T_s})|$ of the ideal controller (thick), with delay and $\beta = 0.34$ (thin) d) $e_i(t)$, slow mobile.

2.2 Measurement Errors and Disturbances

When no measurement noise is present, an intuitive design objective that stems from (7.12) is $S(q) = 0$ and $G(q) = 1$. This would result in perfect disturbance rejection and perfect tracking. However, even if this would have been possible (it is not according to the Bode integral³), it would not be interesting anyway when subject to noise. According to (7.12) this is equivalent to being maximally sensitive to measurement noise. Therefore, it is vital to consider measurement noise in the design and apply measurement filtering if necessary.

Measuring is not an instantaneous procedure, even though the measurements often are considered as samples of a continuous process. This is a relevant approximation in most power control cases. However, some related issues are brought up here.

In WCDMA, measurements are obtained from the fraction δ_s of the slot, which in turn corresponds to $T_s = 1/1500$ s. Typical values (Adachi et al., 1998) of δ_s include $\delta_s = 0.1$ (considering only the four pilot symbols out of

40 symbols) and $\delta_s = 0.25$ (considering the ten first symbols). These values depend on the data rate and channel configuration assigned to the user. A comparison of the filtering effects when considering a full slot average compared to a fractional slot average is found in Fig. 7.7. Aliasing is avoided if the frequency components over the Nyquist frequency are filtered out. This is almost the case when using the local average of the full slot (or measurement period). Conversely, aliasing effects are most likely when adopting local average over fractional slots.

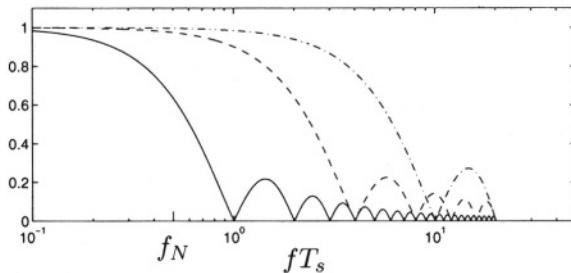


Figure 7.7. Filtering effect with respect to normalized frequency of a local average filter applied to the full slot (solid) and fractional slots $\delta_s = 0.25$ (dashed) and $\delta_s = 0.1$ (dash-dotted). The normalization is with respect to the frequency after down-sampling, which in this case is 40 times smaller.

Approaches for SIR estimation in different systems are discussed in (Ramakrishna et al., 1997; Andersin et al., 1998; Blom et al., 1999; Freris et al., 2001; Kurniawan et al., 2001). It is important to note that especially at low SIR levels, there is a significant bias in the SIR estimation with practically tractable schemes.

2.3 Limited Feedback Bandwidth

To implement a power control algorithm such as the integrating controller in (7.7), information about the error $e_i(t)$ has to be fed back. The feedback signaling bandwidth is limited in real systems. Typically, the communication is restricted to a fixed number of bits per second k . The evident trade-off is between error representation accuracy and feedback command rate. A single bit error representation allows k feedback commands per second, while m_e bits error representation allows k/m_e commands per second. This comparison is further explored in (Gunnarsson, 2001).

Different control error representations are proposed, for example: single bit (the sign of the error) (Salmasi and Gilhousen, 1991), k -bit linear quantizer (Sim et al., 1998) and k -bit logarithmic quantizer (Li et al., 2001). The single bit quantizer optimizes the update rate when the feedback bandwidth is fixed.

$$p_i(t+1) = p_i(t) + \beta \text{sign}(\gamma_i^t(t) - \gamma_i(t)) = p_i(t) + \beta \text{sign}(e_i(t)). \quad (7.14)$$

This is the standardized feedback communication for WCDMA (3GPP, 2003b).

The coded control error is transmitted over the feedback channel, and is therefore subject to possible bit errors. In WCDMA reference cases (Holma and Toskala, 2000) 4% of the single bit commands are assumed erroneous. In one alternative feedback communication scheme for WCDMA (3GPP, 2003b), commands are repeated three consecutive slots. This lowers the update rate to 500 Hz, but improves the command bit error probability from P_b to $P_b^2(3 - 2P_b) << P_b$ (for example from 4 % to 0.47 %).

2.4 Time Delays and Local Stability

Time delays affect stability as with any feedback controlled system, and therefore more careful control actions have to be imposed. However, time delays do not only affect stability. The sensitivity to disturbances is also degraded.

The dynamical behavior of a discrete-time linear system is closely related to the locations of the *closed-loop poles*. These are defined as the roots to the equation of the closed-loop system denominator equal to zero. In case of the integrating controller with closed loop system $G(q)$ as in (7.13), the closed-loop poles are the solutions to

$$q^n(q - 1) + \beta = 0$$

Textbooks on linear systems theory, e.g. (Åström and Wittenmark, 1997), provide the following result for asymptotic stability (i.e. transients decay asymptotically).

Theorem 1 (Asymptotic Stability of Discrete-time Linear Systems) *A time-discrete linear feedback system is uniformly asymptotically stable, if and only if the closed-loop poles are strictly within the unit disc.*

With the integrating controller and in the typical time delay situation $n = 1$, it is easy to verify that this is equivalent to $\beta < 1$. Hence, DPC ($\beta = 1$) is not asymptotically stable when subject to time delays.

Furthermore, the negative effects of the time delay can be reduced by carefully selecting β . It can e.g. be optimized to provide fast reactions ($\beta = 0.34$ for $n = 1$ according to (Gunnarsson et al., 1999)). The time delay also effects the disturbance rejection capabilities of the controller. Figures 7.6c,d illustrate the increased sensitivity to disturbances, and the control error variance for the slow mobile traveling at 2 m/s.

With single-bit error representation as in (7.14) the local loop never converges to fully eliminate the control error. The sign function together with the linear

components combine to result in a triangular wave oscillation of the control error with the period $4n+2$, where n is the time delay (Gunnarsson et al., 2001). This is clearly seen in Figure 7.8c, where we for example identify five oscillation periods between iteration 90 and iteration 120.

Time delays can be handled with careful linear design as discussed above. An alternative is to compensate for the delay in the measurement

$$\gamma_i(t) = p_i(t - n) + \delta_i + g_{ii}(t) - I_i(t). \quad (7.15)$$

If $p_i(t)$ is known in the receiver, the following compensation is plausible

$$\tilde{\gamma}_i(t) = \gamma_i(t) + p_i(t) - p_i(t - n) \quad (7.16)$$

which is referred to as Time Delay Compensation (TDC) (Gunnarsson and Gustafsson, 2001). This internal feedback alters the closed-loop system and is known in control theory as the *Smith predictor* (Åström and Wittenmark, 1997). With TDC, the sensitivity of the all-linear system becomes

$$G(q) = \frac{R(q)}{q^{n_p+n_m}(1+R(q))}, \quad S(q) = \frac{q^{n_p+n_m}}{q^{n_p+n_m}(1+R(q))} \quad (7.17)$$

TDC is particularly interesting when linear design is not an option, perhaps due to standardization as in the WCDMA case. Using (7.14), TDC can be rewritten as

$$\tilde{\gamma}_i(t) = \gamma_i(t) + \beta \sum_{k=t-n}^{t-1} \text{sign}(\gamma_i^k(t) - \gamma_i(t)) \quad (7.18)$$

i.e. adjusting the measurement with issued power control commands that not yet are reflected in the measurement. The benefit of TDC with single-bit error representation is illustrated in Figure 7.8, which also indicates the convergence dependency on β , and the fact that TDC and careful linear design provide roughly the same performance in the linear case.

2.5 Transmitter Power Limitations

One obvious limitation is availability of transmission power. These are partly due to amplifier and battery limitations, but the system may also further restrict the maximum power to limit the mutual interference between connections. It is also important to operate not too close to the limits on average, since the margin is needed to mitigate the fast fading. Furthermore, there is a direct relation between maximum power and coverage of the channel.

2.6 Power Control Error

Some analyses of the CDMA capacity are based on assumptions of the power control error statistics. For example, (Romero-Jerez et al., 2000) discusses ca-

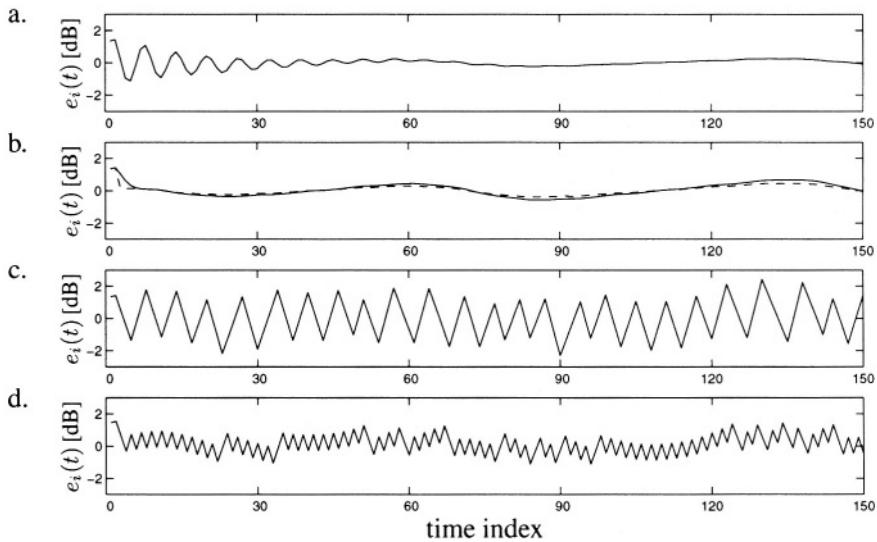


Figure 7.8. Local performance of the algorithms (7.7) in a-b and (7.14) in c-d subject to the typical delay situation $n = 1$. a. $\beta = 0.9$ b. $\beta = 1$ and TDC (dashed) and $\beta = 0.34$ (solid) c. $\beta = 1$ d. $\beta = 1$ and TDC.

pacity using Gaussian power control errors. The fundamental limitations described above in this section effects the error statistics, and simulations here for the integrating controller with single-bit control error representation (7.14) indicates that a Gaussian error model is relevant in most cases.

This algorithm has a superimposed stable oscillation due to time delays and the sign function. As seen in Figure 7.9, the power control error is not Gaussian. However, with both measurement errors and feedback errors as in realistic cases, a Gaussian power control error approximations is clearly plausible.

3. Global Stability

For practical reasons, power control algorithms in cellular radio systems are implemented in a distributed fashion. However, the local loops are interconnected via the interference between the loops, which affects the global dynamics as well as the capacity of the system. An important global issue is whether it is possible to accommodate all users with their service requirements. The power gains reflect the situation from the transmitters to the receivers, and the results are therefore applicable to both the uplink and the downlink.

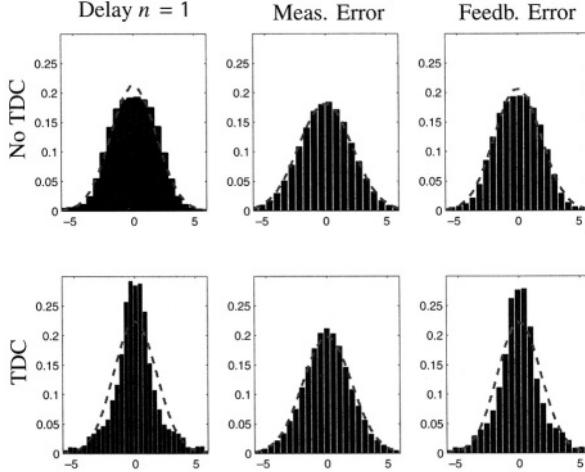


Figure 7.9. Power control error distribution without (top) and with (bottom) time delay compensation when subject to (left) one slot time delay, (middle) 3dB measurement error and (right) 4 % feedback error rate respectively.

3.1 Feasibility

The individual reference SIR:s and the power gains are considered constant in the global level analysis, where the latter is motivated by an assumption that the inner loops perfectly meet the provided SIR reference, and thereby mitigate the fast channel variations:

$$\frac{\bar{\delta}_i \bar{g}_{ii} \bar{p}_i}{\sum_{j \neq i} \bar{g}_{ij} \bar{p}_j + (1 - \bar{\delta}_i) \bar{p}_i \bar{g}_{ii} + \bar{\nu}_i} = \bar{\gamma}_i^t, \quad \forall i \quad (7.19)$$

Note that values in linear scale are used in this section. The aim is to characterize the system load, and a few definitions are needed. Introduce the matrices

$$\bar{\Gamma}_t \triangleq \text{diag}(\bar{\gamma}_1^t, \dots, \bar{\gamma}_m^t), \quad \bar{\mathbf{Z}} = [\bar{z}_{ij}] \triangleq \begin{bmatrix} \bar{g}_{ij} \\ \bar{g}_{ii} \end{bmatrix}, \quad \bar{\Delta} \triangleq \text{diag}(\bar{\delta}_1, \dots, \bar{\delta}_m)$$

and vectors

$$\bar{\mathbf{p}} \triangleq [\bar{p}_i], \quad \bar{\boldsymbol{\eta}} = [\bar{\eta}_i] \triangleq \begin{bmatrix} \bar{\nu}_i \\ \bar{g}_{ii} \end{bmatrix}.$$

The network itself puts restrictions on the achievable SIR's, and there exists an upper limit on the balanced SIR (same SIR to every connection). This is disclosed in the following theorem, neglecting auto-interference and assuming that the noise can be considered zero.

Theorem 2 (Zander, 1992) *With probability one, there exists a unique maximum achievable SIR in the noiseless case*

$$\bar{\gamma}^* = \max\{\bar{\gamma}_0 \mid \exists \bar{\mathbf{p}} \geq 0 : \bar{\gamma}_i \geq \bar{\gamma}_0, \forall i\}.$$

Furthermore, the maximum is given by

$$\bar{\gamma}^* = \frac{1}{\bar{\lambda}^* - 1},$$

where $\bar{\lambda}^*$ is the largest real eigenvalue of $\bar{\mathbf{Z}}$. Note that $\bar{\lambda}^* > 1$ implies that $\bar{\gamma}^* > 0$. Moreover, the optimal power vector $\bar{\mathbf{p}}^*$ is the eigenvector of $\bar{\lambda}^*$ (i.e. $k\bar{\mathbf{p}}^*$ for any $k \in \mathbb{R}^+$ constitutes an optimal power vector).

Considering the noise, the following can be concluded:

Theorem 3 (Zander, 1993) *In the noisy case and with no power limitations, there exist power levels that meet the balanced SIR target $\bar{\gamma}_0^t$ if and only if $\bar{\gamma}_0^t < \bar{\gamma}^*$.*

Consider the following uplink example from Chapter 14⁴

$$\bar{\mathbf{G}}_{UL} = \begin{pmatrix} 0.8 & 0.1 \\ 0.5 & 0.3 \\ 0.2 & 0.6 \end{pmatrix}, \quad \bar{\mathbf{G}} = \begin{pmatrix} 0.8 & 0.5 & 0.2 \\ 0.8 & 0.5 & 0.2 \\ 0.1 & 0.3 & 0.6 \end{pmatrix}, \quad \bar{\mathbf{Z}} = \begin{pmatrix} 1 & 0.63 & 0.25 \\ 1.6 & 1 & 0.4 \\ 0.17 & 0.5 & 1 \end{pmatrix}.$$

where the relation between $\bar{\mathbf{G}}_{UL}$ and $\bar{\mathbf{G}}$ is described in Section 1.2. The optimal balanced SIR from Theorem 2 is thus given by $\bar{\gamma}^* = 0.83$. Hence, only balanced SIR targets less than this imply a feasible power control problem, and it is possible to find power levels that meet the requirements of the users. This is generalized to multiple services below.

The effects of auto-interference are considered in (Godlewski and Nuaymi, 1999; Gunnarsson, 2000), and the requirements in Equation (7.19) can be vectorized to

$$\bar{\mathbf{p}} = \bar{\Gamma}_t \left((\bar{\Delta}^{-1} \bar{\mathbf{Z}} - \mathbf{E}) \bar{\mathbf{p}} + \bar{\Delta}^{-1} \bar{\eta} \right), \quad (7.20)$$

where \mathbf{E} is the identity matrix. Solvability of the equation above is related to feasibility of the related *power control problem*, defined as:

Definition 1 (Feasibility) *A set of target SIR:s $\bar{\Gamma}_t$ is said to be feasible with respect to a network described by $\bar{\mathbf{Z}}$, $\bar{\Delta}$ and $\bar{\eta}$, if it is possible to assign transmitter powers $\bar{\mathbf{p}}$ so that the requirements in Equation (7.20) are met. Analogously, the power control problem ($\bar{\mathbf{Z}}, \bar{\eta}, \bar{\Delta}, \bar{\Gamma}_t$) is said to be feasible under the same condition. Otherwise, the target SIR:s and the power control problem are said to be infeasible.*

The degree of feasibility is described by the *feasibility relative load*, which is defined below.

Definition 2 (Feasibility Relative Load) *Given a power control problem $(\bar{\mathbf{Z}}, \bar{\boldsymbol{\eta}}, \bar{\boldsymbol{\Delta}}, \bar{\boldsymbol{\Gamma}}_t)$, the feasibility relative load $\bar{L}_r \in \mathbb{R}^+$ is defined by*

$$\bar{L}_r = \inf \left\{ \bar{x} \in \mathbb{R} : \frac{1}{\bar{x}} \bar{\boldsymbol{\Gamma}}_t \text{ is feasible} \right\}$$

While Theorems 2 and 3 address the case when all users have the same service, the feasibility relative load describes the situation with multiple services. The following theorem captures the essentials regarding the feasibility relative load.

Theorem 4 (Gunnarsson, 2000) *Given a power control problem $(\bar{\mathbf{Z}}, \bar{\boldsymbol{\eta}}, \bar{\boldsymbol{\Delta}}, \bar{\boldsymbol{\Gamma}}_t)$, the feasibility relative load is obtained as*

$$\bar{L}_r = \max \operatorname{eig} \left\{ \bar{\boldsymbol{\Gamma}}_t (\bar{\boldsymbol{\Delta}}^{-1} \bar{\mathbf{Z}} - \mathbf{E}) \right\}.$$

Moreover, if $\bar{L}_r < 1$, the power control problem is feasible, and there exists an optimal power assignment, given by

$$\bar{\mathbf{p}} = \left(\mathbf{E} - \bar{\boldsymbol{\Gamma}}_t (\bar{\boldsymbol{\Delta}}^{-1} \bar{\mathbf{Z}} - \mathbf{E}) \right)^{-1} \bar{\boldsymbol{\Gamma}}_t \bar{\boldsymbol{\Delta}}^{-1} \bar{\boldsymbol{\eta}}.$$

The power assignment above can of course be seen as a centralized strategy. However, since full information about the network is required to compile $\bar{\mathbf{Z}}$ it is not plausible in practice. The result mainly serves as a performance bound.

Assume perfect receivers in the example, and that the three connections aims at the SIR targets $\bar{\boldsymbol{\Gamma}}_t = \operatorname{diag}(1, 0.43, 0.67)$. According to Theorem 4, the feasibility relative load is $\bar{L}_r = 0.78 < 1$. The power control problem is thus feasible. Note that it is feasible despite the fact that one of the SIR targets is greater than the optimal balanced SIR⁵, $\bar{\gamma}^* = 0.83$.

3.2 Convergence Results

3.2.1 Standard Interference Function. The framework for addressing point-wise convergence of algorithms is based on (Bertsekas and Tsitsiklis, 1989), which has been refined and put into the context of power control algorithms by (Yates, 1995). An underlying assumption is that the G-matrix is constant. Let $\bar{\mathbf{p}}$ denote the controllable resources (in a power control case, it is the power vector). Assume that the requirements to be met can be described by the inequality

$$\bar{\mathbf{p}} \geq \bar{\mathbf{I}}(\bar{\mathbf{p}}), \quad (7.21)$$

where $\bar{\mathbf{I}}(\bar{\mathbf{p}})$ is referred to as the *interference function*. A vector $\tilde{\mathbf{p}} \geq 0$ is a feasible solution if it satisfies (7.21). Moreover, an interference function $\bar{\mathbf{I}}(\bar{\mathbf{p}})$

is feasible if (7.21) has a feasible solution. For an arbitrary interference function, the following definition applies:

Definition 3 (Standard Interference Function) *The interference function is standard if the following properties are satisfied for all $\bar{\mathbf{p}} \geq 0$.*

- Positivity: $\bar{\mathbf{I}}(\bar{\mathbf{p}}) \geq 0$.
- Monotonicity: if $\bar{\mathbf{p}} \geq \bar{\mathbf{p}}'$ then $\bar{\mathbf{I}}(\bar{\mathbf{p}}) \geq \bar{\mathbf{I}}(\bar{\mathbf{p}}')$.
- Scalability: $\alpha \bar{\mathbf{I}}(\bar{\mathbf{p}}) > \bar{\mathbf{I}}(\alpha \bar{\mathbf{p}})$ for all $\alpha > 1$.

Given the requirements in (7.21), we want to study the convergence properties of the class of iterative algorithms described by

$$\bar{\mathbf{p}}(t+1) = \bar{\mathbf{I}}(\bar{\mathbf{p}}(t)). \quad (7.22)$$

When $\bar{\mathbf{I}}(\bar{\mathbf{p}})$ is a standard interference function, the iteration (7.22) is denoted a *standard iterative algorithm*. The corresponding equilibrium is thus

$$\bar{\mathbf{p}}(t+1) = \bar{\mathbf{I}}(\bar{\mathbf{p}}(t)) = \bar{\mathbf{p}}(t).$$

The uniqueness of this equilibrium, is disclosed in the following theorem

Theorem 5 (Yates, 1995) *Let $\bar{\mathbf{I}}(\bar{\mathbf{p}})$ be standard and feasible. If the corresponding standard iterative algorithm (7.22) has an equilibrium, then it is unique.*

Furthermore, the convergence of standard iterative algorithms is provided in Theorem 6

Theorem 6 (Yates, 1995) *If $\bar{\mathbf{I}}(\bar{\mathbf{p}})$ is standard and feasible, then the iterative algorithm (7.22) converges to a unique equilibrium $\bar{\mathbf{p}}_\infty$ for any initial vector $\bar{\mathbf{p}}_0$.*

This techniques can be used to analyze the stability and convergence of the integrating controller in the ideal situation without limitations described in Section 2.

Theorem 7 (Yates, 1995) *The iterative power control algorithm (7.7) converges to a unique equilibrium for any initial power vector if the power control problem is feasible.*

In the ideal case, but with time delays, the following can be concluded

Theorem 8 (Gunnarsson, 2000) *The iterative power control algorithm (7.7) with $\beta = 1$, time delay compensation (7.16) and subject to time delays n converges $n + 1$ times slower than the same algorithm subject to no delays to a unique equilibrium for any initial power vector if the power control problem is feasible.*

3.2.2 Log-Linear Algorithms. The analysis method in the previous section essentially applies to any controller. However, it is difficult to apply when considering algorithms based on log-linear design and when subject to time delays, other than in specific situations.

An alternative method is approximative and based on a linearization of the mutual cross-couplings between the connections via the interference and application of the *small gain theorem* (Zhou et al., 1995). The relative load is introduced in Definition 2. However, to support the stability results in this section, a different definition is also needed

Definition 4 (Cross-coupling Relative Load) *The cross-coupling relative load \bar{L}_{cc} is defined as*

$$\bar{L}_{cc} \triangleq \max_{1 \leq i \leq m} \left(1 - \frac{\bar{\nu}_i}{\bar{I}_i^t} \right)$$

where \bar{I}_i^t is the interference at receiver i at the power control problem equilibrium (essentially using the power vector derived in Theorem 4).

This definition is considering the worst case situation over all receivers. Zero load means no mutual interference in the system, i.e. all the connections operate independently, and a unity load corresponds to an infinite interference, which constitute the upper limit. In the example in Section 3.1, the cross-coupling relative load is $\bar{L}_{cc} = 0.77$.

With this load definition, the following theorem can be concluded.

Theorem 9 (Gunnarsson, 2000) *Let $G(q)$ be the stable closed-loop transfer function of the local loop. Then the global system with linearized cross-couplings is stable if the power control problem is feasible and the following property is satisfied*

$$\sup_{\omega} |G(e^{i\omega})| < 1/\bar{L}_{cc},$$

Note that when cross-couplings can be neglected ($\bar{L}_{cc} = 0$), the power control problem is always feasible, and the only remaining stability requirement is stability of the local loop $G(q)$. Furthermore, a local power control design that meets

$$|G(e^{i\omega})| \leq 1$$

is always stable if locally stable and if the power control problem is feasible. The load level puts no further restrictions on global stability. From (7.17), we conclude that this is the case for the integrating controller (7.7) with TDC (7.16) for any time delay n .

If we consider the integrating controller in (7.7) and the typical time delay situation $n = 1$, the stability requirements in Theorem 9 can be simplified. The

closed-loop system expression in (7.13) yields

$$\sup_w |G(e^{i\omega})| = \begin{cases} \frac{\beta}{\sqrt{(1-\beta)^2(1-\frac{1}{4\beta})}} & \frac{1}{3} \leq \beta < 1 \\ 1 & 0 < \beta < \frac{1}{3} \end{cases} \quad (7.23)$$

The stability requirements in terms of β are illustrated by Figure 7.10. Hence, in highly loaded systems the choice of β is more critical to avoid violating the global stability. In the example $\bar{L}_{cc} = 0.77$ and Theorem 9 can only guarantee

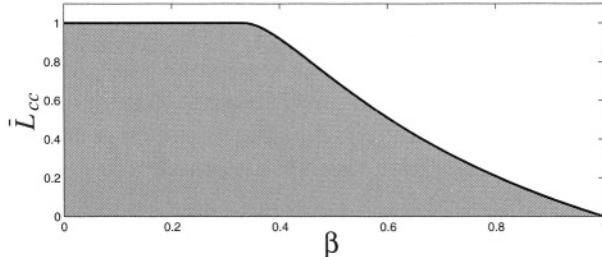


Figure 7.10. Requirements on cross-coupling relative load from Theorem 9 when using an integrating controller, subject to the typical time delay situation $n = 1$.

global stability for $\beta < 0.48$ according to (7.23).

3.2.3 Convergence of the Integrator with Single-Bit Error Representation.

The algorithm described in 7.14 with single-bit error representation never converges to a fixed point. As disclosed in Section 2.4, the relay feedback and the delays cause an oscillatory behavior around the reference signal. Instead, it converges to a region characterized by the following theorem.

Theorem 10 (Herdtnar and Chong, 2000; Gunnarsson, 2000) *If the power control problem is feasible, the algorithm without and with the time delay compensation subject to a delay of totally n samples, $n = 0, 1, 2, \dots$ converges to a region where the SIR error for every connection is bounded (in dB) by*

$$\begin{aligned} \text{Without TDC: } & |\gamma_i^t - \gamma_i(t)| \leq 2(n+1)\beta \\ \text{With TDC: } & |\gamma_i^t - \gamma_i(t)| \leq (n+2)\beta \end{aligned}$$

and β is the step size. The results also hold when subject to auto-interference.

Note that the error bound is tighter when using TDC, and also when using smaller β , which can be interpreted as the step-size. There is thus a trade-off between small tracking errors and fast responses to changes. Furthermore, the importance of well functioning admission and congestion control mechanisms is evident. Power control stability critically rely on their ability to preserve feasibility despite a time-varying network.

3.3 Global Stability Example

Some of the global stability results are illustrated in this section with network simulations using the example from Section 3.1. Figure 7.11 provides SIR error $e_i(t) = \gamma_i^t(t) - \gamma_i(t)$ plots for a number of different parameterizations. Even though the integrating controller (7.7) with perfect error representation and $\beta = 0.9$ yields a locally stable system according to Theorem 1, the global system is unstable and the mobile powers are oscillating between max and min powers. Clearly, the global stability requirements on the closed-loop system are not fulfilled at this load level. Theorem 9 and (7.23) guarantees stability for $\beta = 0.34$ which is illustrated in Figure 7.11b. Almost the same performance is obtained for $\beta = 1$ and time delay compensation in Figure 7.11c – a case which also is globally stable based on Theorem 9. The benefits of TDC are more evident when applied to the integrating controller with single-bit error representation. Comparative performances are provided in Figures 7.11d,e.

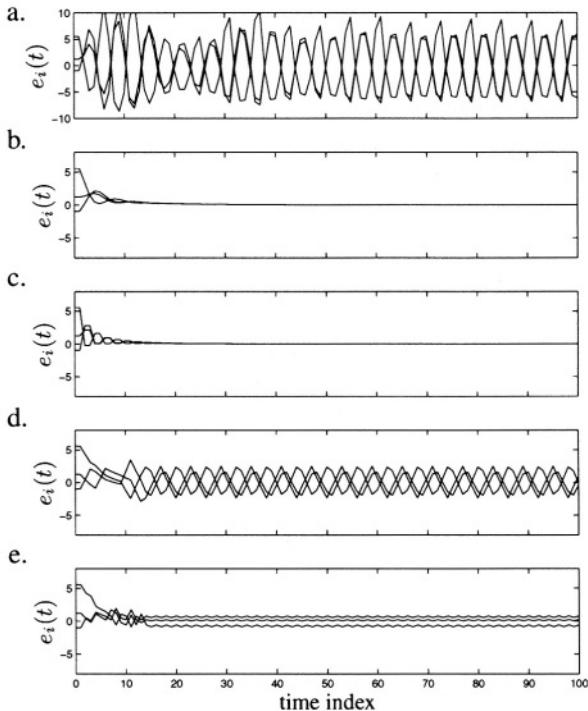


Figure 7.11. Network simulations with the integrating controller (7.7) in a-c and with single-bit error representation (7.14) in d-e. a. $\beta = 0.9$, b. $\beta = 0.34$, c. $\beta = 1$ and TDC, d. $\beta = 1$, e. $\beta = 1$ and TDC.

4. Different Power Control Aspects

Some further power control aspects deserve extra attention, and the objective here is to guide the reader to a wider and more detailed set of references.

4.1 Soft Handover

Hitherto, the models have been focused on a situation, where all mobiles are connected solely to one base station. One core feature in DS-CDMA systems is soft handover, where the mobile can connect to several base stations simultaneously. For best performance, the power is adjusted to compensate for the power gain of the most favorable link. However, since one link is best at the time, the results above are still relevant.

In soft handover, power control becomes significantly different between uplink and downlink. In the uplink, information about the control error to the power control algorithm in the mobile is sent from all connected base stations. It is natural to control the power based on information from the base station with lowest control error, since this is the base station requesting the lowest transmission power. In the case of single-bit error representation, this means to only increase the power if all base stations report a positive control error. However, since the control error information is sent over a feedback channel, unfavorable links cause feedback errors. Therefore, it is important to not only consider the reported control error, but also the feedback channel quality (Grandell and Saloaho, 2001).

In the downlink, all the connected base stations receive control error information from the mobile. Thereby, the relations between the base station powers are maintained. Due to feedback errors, this information might be interpreted differently, and the base station powers might drift apart. To compensate for this drift, a centralized power balancing is proposed in the standards, see (3GPP, 2003c). The following should be seen as replacing equation (7.14)

$$p_i(t+1) = p_i(t) + \beta \text{sign}(e_i(t)) + p_{\text{bal},i} \quad (7.24)$$

The base station computes the average transmission power over a time frame $p_{\text{ave},i}$ and computes the balancing term as

$$p_{\text{bal},i} = (1 - r)(p_{\text{ref},i} - p_{\text{ave},i}), \quad (7.25)$$

where the adjustment ratio r and the reference power $p_{\text{ref},i}$ (possibly individual for each link) are signaled to the base station by the radio network controller (RNC), which controls the base stations.

The need for the central controlling node RNC is evident when considering soft handover. Since the uplink information is available in two different base stations, the actual service quality can only be evaluated in the RNC. Essentially,

the RNC selects the data blocks from the base station with the best SIR (selection combining).

4.2 Outer Loop Power Control

Perceived quality of service might be well correlated to SIR, but it is not possible to set a SIR reference offline, that results in a specific service quality in terms of data rate and block error rate. Furthermore, the SIR estimate could be subject to errors as described in Section 2.2. Therefore, the inner control loop is operated in cascade with an outer loop, which adapts the SIR reference for a specific connection. Typically, the outer loop utilizes information about *block error rate (BLER)* and/or *bit error rate (BER)*.

Reliable communication can be seen as low block error rate requirements, which in turn means that it is very difficult to accurately estimate BLER. The errors appear rather seldom and it takes long time before the BLER estimate is stable. One approach increases the SIR reference significantly when an erroneous block is discovered, and decreases the SIR reference somewhat when an error-free block is received. (Niida et al., 2000) provides experimental results of outer loop power control using this method.

One block comprises many bits. Therefore, it is easier to obtain a good BER estimate. Then the relation between BER and BLER can be utilized to predict BLER based on BER measurements (Kawai et al., 1999).

To support soft handover, the uplink outer loop power control have to be implemented in the node controlling the base stations (RNC in WCDMA). This causes some additional round-trip delay of the control loop.

4.3 Up- and Downlink

The previous sections brought up aspects concerning implementationally differences between uplink and downlink related to soft handover and outer loop power control. Moreover, some other differences are also important.

Power control objectives are naturally different with a low power energy-efficiency focus in the uplink and a desire to utilize all available power in the downlink to use the available resources as much as possible to provide services.

Examples of the latter include High Speed Downlink Shared Channel (HS-DSCH) in the evolved WCDMA standard, where data service users are one by one allocated all base station power except what is needed to support conversational services. This means that the base station powers are always fully utilized, provided that there is always downlink data to transmit. Hence, downlink SIR in (7.5) is given by

$$\bar{\gamma}_i^{\text{DL}}(t) = \frac{\bar{\delta}_i \bar{g}_{ij_i}^{\text{DL}}(t) \bar{P}_{j_i}^{\text{data}}(t)}{\bar{\alpha}_i(\bar{P}^{\max}(t) - \bar{\delta}_i \bar{P}_{j_i}^{\text{data}}(t)) \bar{g}_{ij_i}^{\text{DL}}(t) + \sum_{j \neq j_i} \bar{P}^{\max}(t) \bar{g}_{ij}^{\text{DL}}(t) + \bar{\nu}_i(t)}.$$

where \bar{P}^{\max} is the base station maximum power, and $\bar{P}_{j_i}^{\text{data}}(t)$ the available power for data services at base station j_i . Naturally, the interference situation is more predictable than in the general case, since all base stations ideally are using maximum transmission power. Moreover, the base station resource utilization do not have to be coordinated over the base stations to the same extent for the same reason. This kind of time-sharing of resources and power is analyzed in (Berggren et al., 2001).

The uplink situation is different, since it is not likely that an individual user can utilize all available uplink resources. Instead, several users per base station have to be allocated power and resources for simultaneous use. Uplink power and resource allocation is for example addressed in (Villier et al., 2000). Another difficulty with the uplink is the challenge to characterize the uplink load. This is discussed in Chapter 14.

4.4 Power Control Algorithm Design

This chapter focuses on characteristics and fundamentals of power control algorithms. However, much performance can be gained, at least in theory, by adopting more advanced design methods. Some methods are briefly addressed here, with the objectives to improve link performance in terms of SIR target tracking in Section 4.4.1 and accelerate power control convergence in Section 4.4.2.

4.4.1 Local Power Control Design. The Smith predictor or TDC in (7.16) might compensate for some dynamical effects, but the controllers still show delayed reactions to changes in the power gains. One approach to improve the reactions is to predict the power gain. The following linear model structure is fitted to data

$$g_s(t) = \frac{C(q)}{A(q)} e_s(t), \quad \text{Var} \{e_s^2(t)\} = \sigma_e^2$$

Solve the Diophantine equation

$$q^{m-1}C(q) = A(q)F(q) + G(q)$$

for $F(q)$ and $G(q)$ yields the optimal m -step predictor (Åström and Wittenmark, 1997)

$$\hat{g}_s(t+m|t) = \frac{qG(q)}{C(q)} g_s(t) \quad (7.26)$$

Power gain prediction methods are further studied based on linear model structures (Ericsson and Millnert, 1996; Choel et al., 1999; Ekman et al., 2002) and nonlinear model structures (Ekman and Kubin, 1999; Tanskanen et al.,

1998; Zhang and Li, 1997). It is hard to give a general answer to whether linear or nonlinear models are most appropriate, or whether it is most suitable to predict $g(t)$ or $\tilde{g}(t)$, since it depends on the fading situation and the controller objectives.

When a disturbance model and an optimal predictor is available as above, the step to employ minimum variance control is slightly short (Åström and Wittenmark, 1997; Gunnarsson, 2000):

$$R(q) = \frac{G(q)}{q(q-1)A(q)F(q)}, \quad (7.27)$$

where $F(q)$ and $G(q)$ are obtained from the Diophantine equation above, and precautions are made in order to include integral action into the controller. Minimum-variance controllers are further explored in (Rintamäki and Koivo, 2001; Rintamäki et al., 2002) with linear models of the control error and real-time model estimation and adaptive controller parameters. Controllers like minimum variance control are more aggressive in the control action than the foundational integrating controller. This also means that the stability requirement in Theorem 9 critically limits the system load, which means that the gains in better SIR tracking can be lost due to lower system utilization.

As indicated in previous sections, it is not always justifiable to let a user disturb other connections significantly while aiming at a rather high SIR compared to the propagation conditions. In the proposed algorithm (Almgren et al., 1994), users aiming at using a high power are forced to use a lower SIR. The algorithm expressed in this framework is given by

$$R(q) = \frac{\beta}{q-\beta}, \quad (7.28)$$

which does not include integral action. The algorithm and its convergence is further explored in (Yates et al., 1997).

In real systems, SIR is typically not readily available. One natural idea is to estimate SIR given available measurements. Approaches to do so in different systems are discussed in (Ramakrishna et al., 1997; Andersin et al., 1998; Blom et al., 1999; Freris et al., 2001; Kurniawan et al., 2001). A different approach is to base the power control on the available measurements directly (Ulukus and Yates, 1998).

4.4.2 Accelerating Power Control Convergence. As established in Section 3.2, the foundational integrating power control algorithm converges provided that the power control problem is feasible. The convergence rate of the distributed algorithms from any initial power vector to the equilibrium is also related to the ability of the algorithms to compensate for fast power gain variations. It also depend on the system load (Hanly and Tse, 1999). Methods with the objective to improve the convergence rate, are likely improving

the fading compensation capabilities as well. Such methods typically aims at efficiently and distributedly solving the power control matrix equation (7.20)

$$\bar{\mathbf{p}} = \bar{\Gamma}_t \left((\bar{\Delta}^{-1} \bar{\mathbf{Z}} - \mathbf{E}) \bar{\mathbf{p}} + \bar{\Delta}^{-1} \bar{\boldsymbol{\eta}} \right).$$

This is essentially how the DPC algorithm (7.10) was motivated. Improvements include using successive overrelaxation techniques (Jäntti and Kim, 2000), Steffensen iterations (Li and Gajic, 2002), convex set projections (Rabee et al., 2001), and Gauss-Seidel iterations (Lelic and Gajic, 2002). Since such methods originates from numerical analysis, the robustness against the fundamental limitations in Section 2 needs to be further investigated.

5. Summary

This chapter surveys power control algorithms with a control theoretic perspective. The objective is to adjust the transmitter power $p_i(t)$ to obtain an acceptable signal-to-interference ratio $\gamma_i^t(t) - \gamma_i(t)$. Starting from the foundational integrating controller,

$$p_i(t+1) = p_i(t) + \beta (\gamma_i^t(t) - \gamma_i(t)) = p_i(t) + \beta e_i(t),$$

the practical situation is described via some fundamental limitations including

- limited power level update rate
- measurement errors and disturbances
- limited feedback bandwidth
- time delays and local stability
- transmitter power limitations.

Another natural limitation is global stability – i.e. stability of the entire network when operating several power control algorithms in parallel with mutual interference that needs to be compensated for. Moreover, in practical situations the power control algorithms have to consider connections to several base stations and the fact that it is difficult to address quality of service solely in terms of SIR.

Acknowledgments

This work has been in cooperation with Ericsson Research and supported by VINNOVA, which both are acknowledged.

Notes

1. In some situations it is more convenient to use the related *signal-to-total-interference ratio*, (STIR)

$$\bar{\beta}_i(t) = \frac{\bar{C}_i(t)}{I_i(t) + C_i(t)} = \frac{\bar{\gamma}_i(t)}{1 + \bar{\gamma}_i(t)}.$$

Furthermore, the term SIR/STIR is sometimes replaced by CIR/CTIR (*C* - *carrier*), essentially reflecting the radio frequency situation at the antenna, whereas SIR reflects the base band situation. It can also be important to distinguish between SIR before and after base band processing. Since the absolute values are of no importance here, the choice of terms is irrelevant.

2. For a more rigid discussion on a *q*-operator algebra, the reader is referred to (Åström and Wittenmark, 1997). The intuitive relations to the complex variable *z* of the *z*-transform are also addressed.
 3. A fundamental constraint on the linear control performance and error suppression can be expressed in terms of the Bode integral constraint on *S*.

$$\int_0^\pi \log |S(e^{i\omega})| d\omega = 0$$

This means that it is not possible to obtain $S(e^{i\omega T_s}) = 0$ for all frequencies.

4. Note that the matrices $\tilde{\mathbf{Z}}$ and $\tilde{\mathbf{G}}$ are defined differently in Chapter 14.
 5. The corresponding example with signal-to-total-interference is essentially analogous. The optimal balanced STIR is then $\bar{\beta}^* = 0.45$, and the feasible example is given by $\bar{\beta}^t = (0.5, 0.3, 0.4)$.

References

- 3GPP (2003a). Deployment aspects. Technical Specification TSG RAN 25.943.
 3GPP (2003b). Physical radio procedures. Technical Specification TSG RAN 25.214.
 3GPP (2003c). UTRAN Iub interface NBAP signalling. Technical Specification TSG RAN 25.433.
 Adachi, F., Sawahashi, M., and Suda, H. (1998). Wideband DS-CDMA for next-generation mobile communications systems. *IEEE Communications Magazine*, 36(9).
 Almgren, M., Andersson, H., and Wallstedt, K. (1994). Power control in a cellular system. In *Proc. IEEE Vehicular Technology Conference*, Stockholm, Sweden.
 Andersin, M., Mandayam, N., and Yates, D. (1998). Subspace based estimation of the signal to interference ratio for TDMA cellular systems. *Wireless Networks*, 4(3).
 Åström, K. and Wittenmark, B. (1997). *Computer Controlled Systems – Theory and Design*. Prentice-Hall, Englewood Cliffs, NJ, USA, third edition.
 Berggren, F., Kim, S.-L., Jæntti, R., and Zander, J. (2001). Joint power control and intracell scheduling of DS-CDMA nonreal time data. *IEEE Journal on Selected Areas in Communications*, 19(10).
 Bertsekas, D. P. and Tsitsiklis, J. N. (1989). *Parallel and Distributed Computation*. Prentice Hall, Englewood Cliffs, NJ, USA.

- Blom, J., Gunnarsson, F., and Gustafsson, F. (1998). Constrained power control subject to time delays. In *Proc. International Conference on Telecommunications*, Chalkidiki, Greece.
- Blom, J., Gunnarsson, F., and Gustafsson, F. (1999). Estimation in cellular radio systems. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Phoenix, AZ, USA.
- Chael, S., Chulajata, T., Kwon, H. M., Koh, B.-J., and Hong, S.-C. (1999). Linear prediction at base station for closed loop power control. In *Proc. IEEE Vehicular Technology Conference*, Houston, TX, USA.
- Dietrich, P., Rao, R., Chockalingam, A., and Milstein, L. (1996). A log-linear closed loop power control model. In *Proc. IEEE Vehicular Technology Conference*, Atlanta, GA, USA.
- Ekman, T., Ahlén, A., and Sternad, M. (2002). Unbiased power prediction of Rayleigh fading channels. In *Proc. IEEE Vehicular Technology Conference*, Vancouver, Canada.
- Ekman, T. and Kubin, G. (1999). Nonlinear prediction of mobile radio channels: measurements and MARS model designs. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Phoenix, AZ, USA.
- Ericsson, A. and Millnert, M. (1996). Fast power control to counteract rayleigh fading in cellular radio systems. In *Proc. RVK*, Luleå, Sweden.
- Foschini, G. and Miljanic, Z. (1993). A simple distributed autonomous power control algorithm and its convergence. *IEEE Transactions on Vehicular Technology*, 42(4).
- Freris, N., Jeans, T., and Taaghel, P. (2001). Adaptive SIR estimation in DS-CDMA cellular systems using Kalman filtering. *IEE Electronics Letters*, 37(5).
- Godlewski, P. and Nuaymi, L. (1999). Auto-interference analysis in cellular systems. In *Proc. IEEE Vehicular Technology Conference*, Houston, TX, USA.
- Goldsmith, A. (1997). The capacity of downlink fading channels with variable rate and power. *IEEE Transactions on Vehicular Technology*, 46(3).
- Grandell, J. and Saloaho, O. (2001). Closed-loop power control algorithms in soft handover for WCDMA systems. In *Proc. IEEE International Conference on Communications*, Helsinki, Finland.
- Grandhi, S., Vijayan, R., and Goodman, D. (1994). Distributed power control in cellular radio systems. *IEEE Transactions on Communications*, 42(2).
- Gunnarsson, F. (2000). *Power Control in Cellular Radio System: Analysis, Design and Estimation*. PhD thesis, Linköpings universitet, Linköping, Sweden.
- Gunnarsson, F. (2001). Fundamental limitations of power control in WCDMA. In *Proc. IEEE Vehicular Technology Conference*, Atlantic City, NJ, USA.
- Gunnarsson, F. and Gustafsson, F. (2001). Time delay compensation in power controlled cellular radio systems. *IEEE Communications Letters*, 5(7).

- Gunnarsson, F. and Gustafsson, F. (2002). Power control in cellular radio systems – from a control theory perspective. In *Proc. IFAC World Congress*, Barcelona, Spain.
- Gunnarsson, F., Gustafsson, F., and Blom, J. (1999). Pole placement design of power control algorithms. In *Proc. IEEE Vehicular Technology Conference*, Houston, TX, USA.
- Gunnarsson, F., Gustafsson, F., and Blom, J. (2001). Dynamical effects of time delays and time delay compensation in power controlled DS-CDMA. *IEEE Journal on Selected Areas in Communications*, 19(1).
- Hanly, S. and Tse, D.-N. (1999). Power control and capacity of spread spectrum wireless networks. *Automatica*, 35(12).
- Holma, H. and Toskala, A., editors (2000). *WCDMA for UMTS. Radio Access for Third Generation Mobile Communications*. Wiley, New York, NY, USA.
- Jakes, W. C. (1974). *Microwave mobile communications*. John Wiley & Sons, New York, NY, USA.
- Jäntti, R. and Kim, S.-L. (2000). Second-order power control with asymptotically fast convergence. *IEEE Journal on Selected Areas in Communications*, 18(3).
- Kawai, H., Suda, H., and Adachi, F. (1999). Outer-loop control of target SIR for fast transmit power control in turbo-coded W-CDMA mobile radio. *IEE Electronics Letters*, 35(9).
- Kurniawan, A., Perreau, S., Choi, J., and Lever, K. (2001). SIR-based power control in third generation CDMA systems. In *Proc. International Conference on Information Communications & Signal Processing*, Singapore.
- Lee, T. and Lin, J. (1996). A study on the distributed power control for cellular mobile systems. In *Proc. IEEE Vehicular Technology Conference*, Atlanta, GA, USA.
- Lelic, D. and Gajic, Z. (2002). Gauss-Seidel iterations for SIR-based power updates for 3G wireless CDMA communications networks. In *Proc. Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, USA.
- Li, W., Dubey, V., and Law, C. (2001). A new generic multistep power control algorithm for the LEO satellite channel with high dynamics. To appear in *IEEE Communications Letters*.
- Li, X. and Gajic, Z. (2002). An improved SIR-based power control for CDMA systems using Steffensen iterations. In *Proc. Information Science and Systems Conference*, Princeton, NJ, USA.
- Meyerhoff, H. (1974). Method for computing the optimum power balance in multibeam satellites. *COMSAT Technical Review*, 4(1).
- Niida, S., Suzuki, T., and Takeuchi, Y. (2000). Experimental results of outer-loop transmission power control using wideband-CDMA for IMT-2000. In *Proc. IEEE Vehicular Technology Conference*, Tokyo, Japan.

- Okamura, Y., Ohmori, E., Kawano, T., and Fukuda, K. (1968). Field strength and its variability in VHF and UHF land-mobile radio service. *Review of the Electrical Communication Laboratory*, 16(9-10).
- Rabee, S. A., Sharif, B. S., and Sali, S. (2001). Distributed power control algorithm in cellular radio systems using projection onto convex sets. In *Proc. IEEE Vehicular Technology Conference*, Atlantic City, NJ, USA.
- Ramakrishna, D., Mandayam, N., and Yates, R. (1997). Subspace based estimation of the signal to interference ratio for CDMA cellular systems. In *Proc. IEEE Vehicular Technology Conference*, Phoenix, AZ, USA.
- Rintamäki, M. and Koivo, H. (2001). Adaptive robust power control for WCDMA systems. In *Proc. IEEE Vehicular Technology Conference*, Atlantic City, NJ, USA.
- Rintamäki, M., Zenger, K., and Koivo, H. (2002). Self-tuning adaptive algorithms in the power control of WCDMA systems. In *Proc. Nordic Signal Processing Symposium*, Hurtigruten, Norway.
- Romero-Jerez, J. M., Ruiz-Garcia, M., and Diaz-Estrella, A. (2000). Effects of power control errors and multipath fading on BER in a cellular CDMA system. In *Proc. IEEE Vehicular Technology Conference*, Boston, MA, USA.
- Rosberg, Z. and Zander, J. (1998). Toward a framework for power control in cellular systems. *Wireless Networks*, 4(3).
- Salmasi, A. and Gilhousen, S. (1991). On the system design aspects of code division multiple access (CDMA) applied to digital cellular and personal communications networks. In *Proc. IEEE Vehicular Technology Conference*, New York, NY, USA.
- Shannon, C. (1956). The zero error capacity of a noisy channel. *IRE Transactions On Information Theory*, 2.
- Sim, M., Gunawan, E., Soh, C., and Soong, B. (1998). Characteristics of closed loop power control algorithms for a cellular DS/CDMA system. *IEE Proceedings - Communications*, 147(5).
- Sklar, B. (1997). Rayleigh fading channels in mobile digital communication systems. *IEEE Communications Magazine*, 35(7).
- Sørensen, T. (1998). Correlation model for slow fading in a small urban macro cell. In *Proc. IEEE Personal, Indoor and Mobile Radio Communications*, Boston, MA, USA.
- Tanskanen, J., Mattila, J., Hall, M., Korhonen, T., and Ovaska, S. (1998). Predictive estimators in CDMA closed loop power control. In *Proc. IEEE Vehicular Technology Conference*, Ottawa, Canada.
- Ulukus, S. and Yates, R. (1998). Stochastic power control for cellular radio systems. *IEEE Transactions on Communications*, 46(6).
- Villier, E., Legg, P., and Barrett, S. (2000). Packet data transmissions in a W-CDMA network-examples of uplink scheduling and performance. In *Proc. IEEE Vehicular Technology Conference*, Tokyo, Japan.

- Yates, R. (1995). A framework for uplink power control in cellular radio systems. *IEEE Journal on Selected Areas in Communications*, 13(7).
- Yates, R., Gupta, S., Rose, C., and Sohn, S. (1997). Soft dropping power control. In *Proc. IEEE Vehicular Technology Conference*, Phoenix, AZ, USA.
- Zhang, Y. and Li, D. (1997). Power control based on adaptive prediction in the CDMA/TDD system. In *Proc. IEEE International Conference on Universal Personal Communications*, San Diego, CA, USA.
- Zhou, K., Doyle, J., and K., G. (1995). *Robust and Optimal Control*. Prentice-Hall, Upper Saddle River, NJ, USA.

Chapter 8

AVERAGE OUTAGE DURATION OF WIRELESS COMMUNICATION SYSTEMS*

Lin Yang and Mohamed-Slim Alouini

Department of Electrical and Computer Engineering

University of Minnesota

Minneapolis, MN 55455, USA.

E-mail: <lyang,alouini@ece.umn.edu>

Abstract Outage probability has been traditionally the most commonly used performance measure of wireless communication systems. However, in certain communication system applications such as adaptive transmission schemes, the outage probability does not provide enough information for the overall system design and configuration. In that case, in addition to the outage probability, the frequency of outages (or equivalently the level crossing rate (LCR)) and the average outage duration (AOD) are important performance measures for the proper selection of the transmission symbol rate, interleaver depth, packet length, and/or time slot duration. In this chapter, we present closed-form expressions for the LCR and AOD of some diversity combining schemes of interest with and without multiple co-channel interferences (CCI) and over independent, correlated, and/or unbalanced channels considering minimum signal-to-interference ratio (SIR) and/or desired signal power constraints. More specifically, the chapter presents generic results for the LCR and AOD (i) of maximal ratio combining systems subject to CCI operating over independent identically distributed (i.i.d.) Rician and/or Nakagami fading environments when a minimum desired signal power requirement is specified for satisfactory reception, (ii) of various selection combining diversity scheme in presence of multiple CCI and with both minimum SIR and desired signal power constraints over independent, correlated, and/or unbalanced channels. Corresponding numerical examples and plots illustrating the mathematical formalism are also provided and discussed.

*This work was supported in part by the National Science Foundation Grant No. CCR-9983462.

1. Introduction

In multiuser wireless communication systems which are both interference and power limited, an outage is declared whenever either the output signal-to-interference ratio (SIR) or signal-to-noise ratio falls below its pre-determined threshold. Outage probability has been traditionally the most commonly used important performance measure of cellular mobile radio communication systems (e.g, [1, 2, 3, 4, 5, 6, 7, 8, 9]). In particular, the overall system spectrum efficiency is usually determined by the required downlink outage probability. However, in certain communication system applications such as adaptive transmission schemes, the outage probability does not provide enough information for the overall system design and configuration. In that case, in addition to the outage probability, the frequency of outages (or equivalently the level crossing rate (LCR)) and the average outage duration (AOD) [10, 11, 12, 13] and more generally distribution of the outage duration [14] are important performance measures for the proper selection of the transmission symbol rate, interleaver depth, packet length, and/or time slot duration. LCR and AOD have been introduced and used to reflect the correlation properties (thus the second-order statistics) of fading channels and provide a dynamic representation of the system outage performance.

The performance of wireless systems is severely affected by fading and diversity-combining techniques are often used to mitigate fading and improve the reception quality. These techniques have also been found to be useful in combatting the effects of co-channel interference (CCI) in mobile radio environment. Especially, maximal ratio combining (MRC), which is optimal combining scheme in the absence of CCI but comes at the expense of complexity since MRC requires knowledge of all channel fading parameters. Another extreme case is selection combining (SC), which is one of the simplest micro-diversity combining methods [15], and which always picks the best branch as current operating branch to smooth out the deep fades and severe interferences. Previous related work on the performance analysis of MRC and SC systems has included the following. Outage probability of various diversity combining scheme with/without multiple CCI and over independent, correlated, and/or unbalanced channels considering either/both minimum SIR or/and desired signal power constraints has/have been extensively studied over the past two decades (see for example [1, 2, 3, 4, 5, 6, 16, 17, 18, 19, 7, 8, 9]). On the other hand, recent studies focusing on the analysis of the AOD and the LCR of MRC and SC diversity systems can be found in [20, 21, 22, 23] and in [22, 24, 23], respectively. Previous work on the effect of correlated fading on the AOD and LCR of SC diversity systems was done by

Adachi *et al.* [25]. More specifically, [25] presented a general method for calculating the LCR of correlated dual-branch SC system without CCI but ended up with approximate expressions for both the LCR and AOD.

All above published analytical work on the LCR and AOD of cellular mobile radio systems is still limited to purely power-limited systems [25, 20, 21, 22, 24, 23] or interference-limited systems subject to Rayleigh type of fading [10, 12]. Although the pure scattering (Rayleigh) fading model is reasonable for medium or large macrocellular systems, a line-of-sight or specular paths are more likely to exist in microcellular or picocellular indoor environments because cell size is smaller. In such kind of environments, the fading tends to rather follow a Rician or Nakagami type of distribution. Furthermore, in previous work on AOD evaluation, it was assumed that an outage is declared if either the desired signal power or the SIR falls below a predetermined protection ratio. While, this criterion is acceptable for purely power-limited or interference-limited systems, it tends to under-estimate the outage in particular for sectorized or lightly loaded systems with CCI in which an outage can occur not only because of low SIR but also because deep fades will result in a desired signal power below a minimum required level for satisfactory detection. To fill the remaining void in the area of performance analysis of LCR and AOD of diversity systems, the authors recently studied the LCR and AOD of some diversity combining schemes of interest with and without multiple CCI and over independent, correlated, and/or unbalanced channels considering minimum SIR and/or desired signal power constraints [13, 26, 27]. This chapter presents the main results of this study without going into all the details of the derivations.

2. Impact of CCI on the LCR and AOD of MRC

While special attention has been paid to the impact of the fading statistics of the desired and interfering users on the system outage probability (see for example [1, 3, 5, 28, 29, 7, 30] and references therein), published analytical work on the LCR and AOD of MRC-like cellular mobile radio systems is still limited to interference-limited systems subject to Rayleigh type of fading [10, 12]. In this section, we present closed-form expressions for the LCR and AOD of MRC in generalized fading environments and with the consideration of both minimum SIR and minimum desired signal power requirements for satisfactory reception.

2.1 Desired and Interference Signals Model

We consider a cellular mobile radio system in which the desired and interfering signals are independent and may have different fading statistics [31, 5, 32, 33, 7]. The desired signal is received over L independent identically distributed (i.i.d.) diversity (time, frequency, or space) paths with the same average fading power Ω_D and subject to flat Rayleigh, Nakagami (with fading parameter m_D), or Rician (with Rician factor K_D) type of fading. The total N_I active co-channel interfering signals (over the L diversity paths) are assumed to be i.i.d. with the same average fading power Ω_I and subject to flat Rayleigh, Nakagami (with fading parameter m_I), or Rician (with Rician factor K_I) type of fading.

The output SIR λ for an interference-limited system with MRC is given by [28, 34]

$$\lambda = \frac{\alpha_D^2}{\alpha_I^2} = \frac{\sum_{l=1}^L \alpha_{D,l}^2}{\sum_{n=1}^{N_I} \alpha_{I,n}^2}, \quad (8.1)$$

where α_D and α_I are the composite amplitudes of the desired user and co-channel interferers at the combiner output, $\alpha_{D,l}$ is the fading amplitude of the desired user over the l th diversity path and $\alpha_{I,n}$ is the fading amplitude of the n th co-channel interfere.

2.2 Method of Analysis

We consider wireless communication systems in which multiple CCI are present. Depending on propagation and fading conditions affecting the desired and interfering signals, the system is alternatively interference-limited or power-limited. For such kind of systems, an overall outage is declared whenever either the output SIR λ falls below a pre-determined threshold λ_{th} , i.e., $\lambda = \frac{\alpha_D^2}{\alpha_I^2} < \lambda_{\text{th}}$, or the received signal power of the desired user $s_D = \alpha_D^2$ falls below another pre-determined threshold s_{th} , i.e., $s_D < s_{\text{th}}$. In this section, we seek to evaluate the AOD, $T(\lambda_{\text{th}}, s_{\text{th}})$ [seconds] which is a measure to describe how long in average the system remains in the outage status.

Mathematically speaking, it can be shown that $T(\lambda_{\text{th}}, s_{\text{th}})$ is given by [13, Eqn. 8]

$$T(\lambda_{\text{th}}, s_{\text{th}}) = \frac{P_{\text{out}}}{N(\lambda_{\text{th}}, s_{\text{th}})}, \quad (8.2)$$

where $N(\lambda_{\text{th}}, s_{\text{th}})$ is the joint LCR that the SIR process crosses λ_{th} or the desired signal power process crosses s_{th} and $P_{\text{out}} = P[\lambda < \lambda_{\text{th}} \text{ or } s_D <$

s_{th}] is the system outage probability when a minimum desired signal power constraint is present.

The outage probability for various desired/interfering signal fading scenarios with a minimal signal power requirement has been extensively studied and closed-form expressions are available for several fading scenario combinations of interest (see for example [1, 3, 5, 7] for Nakagami/Rician¹, Rician/Nakagami, and Nakagami/Nakagami fading combinations and no diversity reception). Therefore, only the LCR $N(\lambda_{\text{th}}, s_{\text{th}})$ is needed to obtain the AOD in all above cases. However, it is worthwhile to point out that our approach and resulting new expressions for the LCR are also applicable to the diversity systems employing MRC.

Conditioning on α_I and using the total probability theorem, it can be shown that the LCR with a minimum desired signal power requirement can be written as

$$\begin{aligned} N(\lambda_{\text{th}}, s_{\text{th}}) = & N\left(\alpha_D = \sqrt{\lambda_{\text{th}}} \alpha_I \mid \alpha_I \geq \sqrt{s_{\text{th}}/\lambda_{\text{th}}}\right) P\left(\alpha_I \geq \sqrt{s_{\text{th}}/\lambda_{\text{th}}}\right) \\ & + N\left(\alpha_D = \sqrt{s_{\text{th}}}\right) P\left(\alpha_I < \sqrt{s_{\text{th}}/\lambda_{\text{th}}}\right), \end{aligned} \quad (8.3)$$

where $N(\alpha_D = \sqrt{s_{\text{th}}})$ is the LCR of the signal power α_D^2 at level s_{th} and $N\left(\alpha_D = \sqrt{\lambda_{\text{th}}} \alpha_I \mid \alpha_I \geq \sqrt{s_{\text{th}}/\lambda_{\text{th}}}\right)$ is the LCR of the SIR λ at level λ_{th} given $\alpha_I \geq \sqrt{s_{\text{th}}/\lambda_{\text{th}}}$.

Closed-form expressions for $P\left(\alpha_I < \sqrt{s_{\text{th}}/\lambda_{\text{th}}}\right)$, $N(\alpha_D = \sqrt{s_{\text{th}}})$, $P\left(\alpha_I \geq \sqrt{s_{\text{th}}/\lambda_{\text{th}}}\right)$ are available in the literature (e.g [35, 20, 36]), and $N\left(\alpha_D = \sqrt{\lambda_{\text{th}}} \alpha_I \mid \alpha_I \geq \sqrt{s_{\text{th}}/\lambda_{\text{th}}}\right)$ can be obtained by modifying the formula provided in [36, Eqn. (2.90)] as

$$N\left(\alpha_D = \sqrt{\lambda_{\text{th}}} \alpha_I \mid \alpha_I \geq \sqrt{s_{\text{th}}/\lambda_{\text{th}}}\right) = \int_0^\infty \dot{r} f_{r, \dot{r}|A}\left(\sqrt{\lambda_{\text{th}}}, \dot{r}\right) d\dot{r}, \quad (8.4)$$

where $r = \alpha_D/\alpha_I$, A denotes the event $\{\alpha_I \geq \sqrt{s_{\text{th}}/\lambda_{\text{th}}}\}$, and $f_{r, \dot{r}|A}(r, \dot{r})$ is the conditional joint probability density function (JPDF) of r and \dot{r} given $\alpha_I \geq \sqrt{s_{\text{th}}/\lambda_{\text{th}}}$. As an extension of the formula [37, Eqn. (9)] (or equivalently [10, Eqn. (15)]), it can be shown that $f_{r, \dot{r}|A}(r, \dot{r})$ is given by

$$\begin{aligned} f_{r, \dot{r}|A}(r, \dot{r}) = & \int_{\sqrt{\frac{s_{\text{th}}}{\lambda_{\text{th}}}}}^\infty \int_{-\infty}^\infty \alpha_I^2 f_{\alpha_D}(\alpha_I r) f_{\dot{\alpha}_D}(\dot{r} \alpha_I + \alpha_I \dot{r}) \\ & \times f_{\alpha_I|A}(\alpha_I) f_{\dot{\alpha}_I}(\dot{\alpha}_I) d\dot{\alpha}_I d\alpha_I, \end{aligned} \quad (8.5)$$

where $f_{\alpha_I|A}(\alpha_I)$ denote the truncated probability density function (PDF) of α_I given $\alpha_I \geq \sqrt{s_{\text{th}}/\lambda_{\text{th}}}$ and which can be obtained as

$$f_{\alpha_I|A}(\alpha_I) = \frac{f_{\alpha_I}(\alpha_I)}{P(\alpha_I \geq \sqrt{s_{\text{th}}/\lambda_{\text{th}}})}, \quad \alpha_I \geq \sqrt{\frac{s_{\text{th}}}{\lambda_{\text{th}}}}. \quad (8.6)$$

2.3 Application to Common Fading Scenarios

In this section, we provide closed-form expressions for the LCR corresponding to typical fading combinations of interest: Nakagami/Nakagami, Rician/Nakagami, and Nakagami/Rician.

2.3.1 Nakagami/Nakagami Scenario. We consider the case where both desired user and co-channel interferers are subject to Nakagami fading. It can be shown after considerable manipulations that the LCR with a minimum power requirement $N(\lambda_{\text{th}}, s_{\text{th}})$ is given by

$$\begin{aligned} N(\lambda_{\text{th}}, s_{\text{th}}) &= \sqrt{2\pi} f_D \frac{S_n^{\frac{2Lm_D-1}{2}}}{\Gamma(Lm_D)} \exp(-S_n) \left[1 - \frac{\Gamma(N_I m_I, \lambda_{nn} S_n)}{\Gamma(N_I m_I)} \right] \\ &\quad + \frac{\sqrt{2\pi} \Gamma(Lm_D + N_I m_I - \frac{1}{2}, S_n(1 + \lambda_{nn}))}{\Gamma(N_I m_I) \Gamma(Lm_D)} \\ &\quad \times \left(f_D^2 + \frac{f_I^2}{\lambda_{nn}} \right)^{\frac{1}{2}} \frac{\lambda_{nn}^{N_I m_I}}{(1 + \lambda_{nn})^{Lm_D + N_I m_I - \frac{1}{2}}}, \end{aligned} \quad (8.7)$$

where $S_n = \frac{s_{\text{th}} m_D}{\Omega_D}$, $\lambda_{nn} = \frac{\Omega_D m_I}{\lambda_{\text{th}} \Omega_I m_D}$, f_D and f_I are the maximum Doppler frequency shifts for the desired and interfering signals, respectively, and $\Gamma(\cdot, \cdot)$ is the incomplete Gamma function [38].

Special Cases

- Interference-limited case

The LCR $N(\lambda_{\text{th}})$ for interference-limited systems can be obtained by setting $s_{\text{th}} = 0$ in (8.7) which results into

$$\begin{aligned} N(\lambda_{\text{th}}) &= \frac{\sqrt{2\pi} \Gamma(Lm_D + N_I m_I - \frac{1}{2})}{\Gamma(N_I m_I) \Gamma(Lm_D)} \left(f_D^2 + \frac{f_I^2}{\lambda_{nn}} \right)^{\frac{1}{2}} \\ &\quad \times \frac{\lambda_{nn}^{N_I m_I}}{(1 + \lambda_{nn})^{Lm_D + N_I m_I - \frac{1}{2}}}. \end{aligned} \quad (8.8)$$

If $f_D = f_I = f_m$ and $m_D = m_I = 1$, (8.8) further reduces to the LCR of the interference-limited system in the Rayleigh/Rayleigh

fading environment, which is given by

$$N(\lambda_{\text{th}}) = \frac{\sqrt{2\pi} f_m \Gamma(N_I + \frac{1}{2})}{\Gamma(N_I)} \left(\frac{1}{1 + \frac{\lambda_{\text{th}} \Omega_I}{\Omega_D}} \right)^{N_I} \sqrt{\frac{\lambda_{\text{th}} \Omega_I}{\Omega_D}}, \quad (8.9)$$

in agreement with [10, Eqn. (17)] when $L = 1$, as expected.

- Power-limited case

Setting $\lambda_{\text{th}} = 0$ in (8.7), we obtain the LCR $N(s_{\text{th}})$ of the signal power s_D at level s_{th} (purely power-limited case) which is given by

$$N(s_{\text{th}}) = \sqrt{2\pi} f_D \frac{S_n^{\frac{2Lm_D-1}{2}}}{\Gamma(Lm_D)} \exp(-S_n), \quad (8.10)$$

in agreement with Yacoub *et al.* L -fold MRC diversity result over i.i.d. Nakagami paths [22, Eqn. (36)] (or equivalently their result for no diversity case given in [35, Eqn. (17)]).

2.3.2 Rician/Nakagami Scenario. When the desired user is subject to Rician type of fading whereas the co-channel interferers are subject to i.i.d. Nakagami fading, it can be shown that the LCR with a minimum power requirement is given by

$$\begin{aligned} N(\lambda_{\text{th}}, s_{\text{th}}) &= \sqrt{2\pi} f_D \frac{S_r^{\frac{L}{2}}}{(LK_D)^{\frac{L-1}{2}}} \exp(-LK_D - S_r) \\ &\times I_{L-1} \left(2\sqrt{LK_D S_r} \right) \left[1 - \frac{\Gamma(N_I m_I, \lambda_{rn} S_r)}{\Gamma(N_I m_I)} \right] \\ &+ \frac{2^{\frac{3-L-2N_I m_I}{2}} \sqrt{\pi}}{\Gamma(N_I m_I) (LK_D)^{\frac{L-1}{2}}} \frac{\lambda_{rn}^{N_I m_I}}{(1 + \lambda_{rn})^{\frac{L}{2} + N_I m_I}} \\ &\times Q_{L+2N_I m_I-1, L-1} \left(\sqrt{\frac{2LK_D}{1 + \lambda_{rn}}}, \sqrt{2S_r(1 + \lambda_{rn})} \right) \\ &\times \exp \left(-\frac{LK_D \lambda_{rn}}{1 + \lambda_{rn}} \right) \left(f_D^2 + \frac{f_I^2}{\lambda_{rn}} \right)^{\frac{1}{2}}, \end{aligned} \quad (8.11)$$

where $S_r = \frac{s_{\text{th}}(K_D+1)}{\Omega_D}$, $\lambda_{rn} = \frac{\Omega_D m_I}{\lambda_{\text{th}} \Omega_I (K_D+1)}$, and $Q_{m,n}(\cdot, \cdot)$ is the Nuttall Q -function defined by [39]. Similarly, one can get the closed-form results of the LCR for purely interference-limited/power-limited system in this scenario by setting $s_{\text{th}} = 0/\lambda_{\text{th}} = 0$ in (8.11) [13].

2.3.3 Nakagami/Rician Scenario. The LCR can be shown to be given in this scenario by

$$\begin{aligned}
 N(\lambda_{\text{th}}, s_{\text{th}}) = & \sqrt{2\pi} f_D \frac{S_n^{\frac{2Lm_D-1}{2}}}{\Gamma(Lm_D)} \exp(-S_n) \left[1 - Q_{N_I} \left(\sqrt{2N_I K_I}, \sqrt{2S_n \lambda_{nr}} \right) \right] \\
 & + \frac{2^{\frac{3-N_I-2Lm_D}{2}} \sqrt{\pi}}{\Gamma(Lm_D) (N_I K_I)^{\frac{N_I-1}{2}}} \frac{\lambda_{nr}^{\frac{N_I+1}{2}}}{(1 + \lambda_{nr})^{\frac{N_I}{2} + Lm_D}} \\
 & \times Q_{N_I+2Lm_D-1, N_I-1} \left(\sqrt{\frac{2N_I K_I \lambda_{nr}}{1 + \lambda_{nr}}}, \sqrt{2S_n(1 + \lambda_{nr})} \right) \\
 & \times \exp \left(-\frac{N_I K_I}{1 + \lambda_{nr}} \right) \left(f_D^2 + \frac{f_I^2}{\lambda_{nr}} \right)^{\frac{1}{2}}, \tag{8.12}
 \end{aligned}$$

where $\lambda_{nr} = \frac{\Omega_D(K_I+1)}{\lambda_{\text{th}} \Omega_I m_D}$ and $Q_n(a, b) = Q_{n,n-1}(a, b)/a^{n-1}$ is the generalized Marcum Q -function [39]. Similarly, the closed-form results of the LCR for purely interference-limited/power-limited system in this scenario can be obtained by setting $s_{\text{th}} = 0/\lambda_{\text{th}} = 0$ in (8.11) [13].

2.4 Numerical Examples

Figure 8.1 plots the AOD versus the normalized SIR threshold $\frac{\Omega_D}{\Omega_I \lambda_{\text{th}}}$ when $L = 1$, $N_I = 6$, $m_D = 2$, $m_I = 2$, and for various values of the normalized signal power thresholds $\frac{s_{\text{th}}}{\Omega_D}$. From this figure, we can see that the minimum desired signal power requirement causes a floor on the AOD, since for $s_{\text{th}} > 0$ increasing the normalized SIR threshold above a particular value does not reduce the AOD and the system becomes power limited. This floor and the value of the normalized SIR threshold at which this floor occurs depends on s_{th}/Ω_D for a fixed fading parameters. This can be seen from Fig. 8.2 which shows the AOD versus the normalized desired signal power threshold for various values of the normalized SIR threshold. On the other hand, Figs. 8.3 and 8.4 show the effects of the speeds of desired and interfering users on the AOD. Figure 8.3 plots the AOD versus the normalized SIR threshold $\frac{\Omega_D}{\Omega_I \lambda_{\text{th}}}$ when $L = 1$, $N_I = 6$, $f_D = f_I = 40$ Hz, $\frac{s_{\text{th}}}{\Omega_D} = -20$ dB, $m_D = 2$, and $m_I = 2$, and for various values of the desired user maximum Doppler frequencies. It is clear that the desired user Doppler frequency (or equivalently speed) has a significant impact on the system AOD. And from Fig. 8.4, we can also conclude that the interfering users speed has a negligible effect of the system AOD especially in the medium to high normalized SIR threshold range. Furthermore, Figure 8.5 shows that for interference-limited system a relatively significant decrease in the AOD

is obtained as the number of diversity paths L increases and on the other hand the increase in the number of interferers N_I results in performance degradation in AOD, especially when normalized SIR threshold $\frac{\Omega_D}{\Omega_I \lambda_{th}}$ is small. Obviously diminishing diversity gain is obtained as the number of diversity paths L increases.

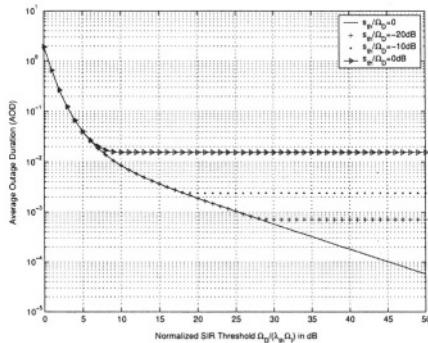


Figure 8.1. AOD versus $\frac{\Omega_D}{\Omega_I \lambda_{th}}$ for various values of $\frac{s_{th}}{\Omega_D}$ when $L = 1$, $N_I = 6$, $f_D = f_I = 40$ Hz, $m_D = 2$, and $m_I = 2$.

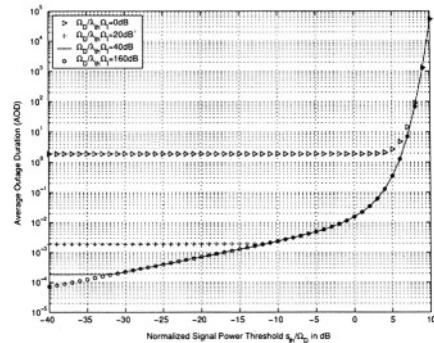


Figure 8.2. AOD versus $\frac{s_{th}}{\Omega_D}$ for various values of $\frac{\Omega_D}{\Omega_I \lambda_{th}}$ when $L = 1$, $N_I = 6$, $f_D = f_I = 40$ Hz, $m_D = 2$, and $m_I = 2$.

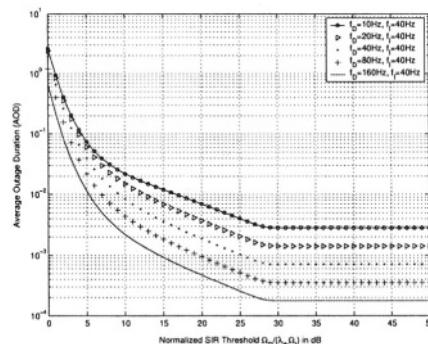


Figure 8.3. AOD versus $\frac{\Omega_D}{\Omega_I \lambda_{th}}$ for various values of the f_D when $L = 1$, $N_I = 6$, $f_I = 40$ Hz, $\frac{s_{th}}{\Omega_D} = -20$ dB, $m_D = 2$, and $m_I = 2$.

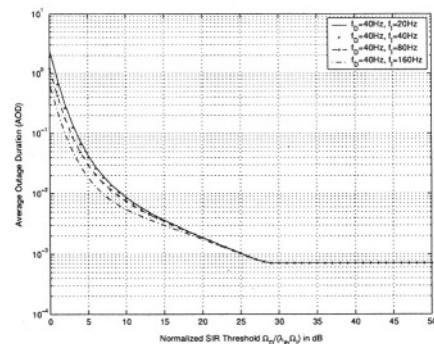


Figure 8.4. AOD versus $\frac{\Omega_D}{\Omega_I \lambda_{th}}$ for various values of f_I when $L = 1$, $N_I = 6$, $f_D = 40$ Hz, $\frac{s_{th}}{\Omega_D} = -20$ dB, $m_D = 2$, and $m_I = 2$.

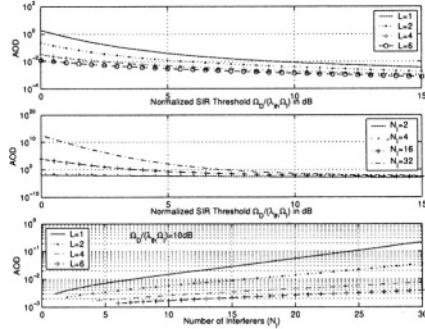


Figure 8.5. Effects of N_I and L on aod when $L = 1$, $N_I = 6$, $m_D = m_I = 2$, and $f_D = f_I = 40\text{Hz}$.

3. Impact of CCI and Fading Correlation on the LCR and AOD of SC

3.1 System Model

For generality purpose, we consider a cellular mobile radio system where the mobile station has an L -branch antenna SC diversity combiner. The received signal at the i th antenna, operating in presence of N_I co-channel interferers, may be written as

$$r_i(t) = \alpha_{D,i} \cos(\omega_c t + \theta_i(t)) + \sum_{j=1}^{N_I} \alpha_{I,ij} \cos(\omega_c t + \psi_{ij}(t)) + n_i(t), \quad (8.13)$$

where ω_c is the carrier frequency, $\theta_i(t)$ and $\psi_{ij}(t)$ are the random phases of the desired signal and j th interfering signal, respectively, $n_i(t)$ is the additive Gaussian noise, $\alpha_{D,i}$ and $\alpha_{I,ij}$ are the received random envelopes at any time t of desired signal and j th interfering signal, respectively.

We consider the system in which the desired and interfering signals are independent and may have different fading statistics [31, 5, 32, 33, 7]. The received desired signal envelopes $\alpha_{D,i}, i = 1, \dots, L$, are assumed to be independent but not necessarily identically distributed with the average fading power $\Omega_{D,i}, i = 1, \dots, L$ and subject to flat Rayleigh, Nakagami, or Rician type of fading. The total $N_I L$ received co-channel interfering signal envelopes $\alpha_{I,ij}, i = 1, \dots, L, j = 1, \dots, N_I$, over the L diversity paths are assumed to be i.i.d. with the same average fading power Ω_I and subject to flat Rayleigh, Nakagami, or Rician type of fading.

SC is one of the simplest micro-diversity combining methods [15], which always picks the best branch as current operating branch to smooth out the deep fades and severe interferences. When a SC combiner is subject to CCI, the selection algorithm is not unique [40]. More specifically, the selection of branches for SC can be based on total (desired plus interference) power (i.e. best S+I algorithm), desired signal power (best signal power algorithm or best S algorithm), or SIR of the received signal at each branch output (best SIR algorithm).

Of the three algorithms, the best SIR algorithm usually provides the best performance for interference-limited systems, but is the most complex to implement, the best signal power algorithm requires identification and separation of desired signal and interferers (which can be achieved using different pilots [40]), and the best S+I algorithm is the most practical to implement. When SC is employed at the antenna array, the output SIR λ depends on the selection algorithm and as such is given by

$$\lambda = \frac{\alpha_D^2}{\alpha_I^2} = \begin{cases} \frac{\max_l \{\alpha_{D,l}^2\}}{\sum_{j=1}^{N_I} \alpha_{I,j}^2}, & \text{For Best S Alg.} \\ \max_l \{\lambda_l\} = \max_l \left\{ \frac{\alpha_{D,l}^2}{\sum_{j=1}^{N_I} \alpha_{I,j}^2} \right\}, & \text{For Best SIR Alg. , (8.14)} \\ \frac{\alpha_{D,i}^2}{\sum_{j=1}^{N_I} \alpha_{I,j}^2} \Big|_{\alpha_{D,i}^2 \rightarrow \sum_{j=1}^{N_I} \alpha_{I,j}^2 \max_l}, & \text{For Best S+I Alg.} \end{cases}$$

where α_D and α_I are the amplitude of the desired user and the composite amplitude of the co-channel interferers both at the SC combiner output.

3.2 Performance Comparison of Different SC Algorithms in Presence of CCI

Previous related work on the performance analysis of SC has included the following. Outage probability of SC in presence of CCI but without a minimum signal power requirement has been extensively studied over the past two decades (see for example [1, 2, 3, 4, 5, 6, 7, 8, 9]). Specifically, Jakes [40] derived the outage probability for the interference-limited case of SC in a single interferer Rayleigh fading scenario using the best signal power algorithm. Based on the best S+I algorithm, Schiff [41] presented the PDF of output SIR of SC in a single interferer Rayleigh fading scenario. Sowerby and Williamson [2] considered the multiple interferers case with SC over Rayleigh fading using both best SIR and best S+I algorithms. Abu-Dayya and Beaulieu [6] generalized the analysis of the interference-limited case of SC to the Nakagami fading environment for all three algorithm mentioned above (but obtained no closed-form for best S+I algorithm case). Finally, Okui considered dual-branch SC in

correlated Nakagami fading using best signal power algorithm in [4] and using best SIR algorithm in [8]. Yang and Alouini in [9] recently studied the outage probability of SC in single interferer Rayleigh fading environment for best signal power algorithm with the consideration of both minimum SIR and signal power constraints. On the other hand, the LCR and AOD of SC over various fading channels but in absence of interference has been recently investigated in [24, 22, 23]. In [26], we complement these previous studies and present uniform analytical approaches for the outage probability, LCR, and AOD computation of SC diversity schemes in presence of multiple CCI. In our analysis we consider both minimum SIR and desired signal power constraints for all three algorithms. We then specialize our analysis to a low-complexity dual-branch SC receiver subject to multiple interferers over Rayleigh fading channels. This particular interference/fading scenario applies to small-size mobile terminals operating in heavily loaded macrocellular networks. Although more general fading scenarios (such as Rice and Nakagami) are also of interest, the relatively simple closed-form results obtained for the special case of interest provide useful insight for the design of SC diversity systems in presence of CCI.

3.2.1 General Method for SC Performance Analysis. For communication systems which are both interference and power limited, the outage is defined in section 2.2. Extending the method used by Schiff in [41], outage probability P_{out} can be written for all three selection algorithms of SC as

$$\begin{aligned}
 P_{\text{out}} &= 1 - \Pr[\text{System is not in outage}] \\
 &= 1 - \sum_{i=1}^L \Pr(\lambda_i > \lambda_{\text{th}}, s_{D,i} > s_{\text{th}}, \rho_i \text{ is the max of } \{\rho_k\}_{k=1}^L) \\
 &= 1 - \sum_{i=1}^L \underbrace{\int \int \int}_{s_D, \lambda, \rho \in G} f_i(s_D, \lambda, \rho) \prod_{\substack{k=1 \\ k \neq i}}^L F_k(\rho) d\rho d\lambda ds_D,
 \end{aligned} \tag{8.15}$$

where $f_i(s_D, \lambda, \rho)$ is the JPDF of $s_{D,i}$, λ_i , and ρ_i at the i th branch, $F_k(\rho)$ is the CDF of ρ_k at the k th branch, G stands for the non-outage area with respect to s_D , λ , and ρ , the symbol ρ represents the statistics used for selection algorithm. When the best signal power algorithm is employed, $\rho = s_D$ and $G = \{\lambda \geq \lambda_{\text{th}} \cap s_D \geq s_{\text{th}}\}$. If the selection is based on the best SIR algorithm, then $\rho = \lambda$ and still $G = \{\lambda \geq \lambda_{\text{th}} \cap s_D \geq s_{\text{th}}\}$. When the best S+I algorithm is employed, we have $\rho = s_D + s_I = \alpha_D^2 + \alpha_I^2$ and $G = \{s_D \geq s_{\text{th}} \cap s_D \leq \rho \leq \frac{1+\lambda_{\text{th}}}{\lambda_{\text{th}}} s_D\}$ since the non-outage condition

$\{\lambda = \frac{s_D}{s_I} = \frac{s_D}{\rho - s_D} \geq \lambda_{\text{th}} \text{ and } s_D \geq s_{\text{th}}\}$ is equivalent to $\{s_D \leq \rho \leq \frac{1+\lambda_{\text{th}}}{\lambda_{\text{th}}} s_D \text{ and } s_D \geq s_{\text{th}}\}$.

To evaluate the AOD, $T(\lambda_{\text{th}}, s_{\text{th}})$ in (8.2) for SC system, we also need to calculate LCR $N(\lambda_{\text{th}}, s_{\text{th}})$ which can be obtain by using the method provided in section 2.2, except that the conditional JPDF $f_{r,\dot{r}|A}(r, \dot{r})$ for SC can be written as [26, Eqn. (20)]

$$f_{r,\dot{r}|A}(r, \dot{r}) = \int_{\sqrt{\frac{s_{\text{th}}}{\lambda_{\text{th}}}}}^{\infty} \int_{-\infty}^{\infty} \alpha_I^2 f_{\alpha_D, \dot{\alpha}_D}(\alpha_I r, \dot{r} \alpha_I + \dot{\alpha}_I r) \\ \times f_{\alpha_I|A}(\alpha_I) f_{\dot{\alpha}_I}(\dot{\alpha}_I) d\dot{\alpha}_I d\alpha_I, \quad (8.16)$$

where $f_{\alpha_D, \dot{\alpha}_D}(\alpha, \dot{\alpha})$ is the JPDF of α_D and its time derivative $\dot{\alpha}_D$ which depends on the selection algorithm employed in SC system and can be obtained in a uniform way for all three algorithms as [26, Eqn. (16)]

$$f_{\alpha_D, \dot{\alpha}_D}(\alpha, \dot{\alpha}) = \sum_{i=1}^L f_{\dot{\alpha}_{D,i}}(\dot{\alpha}) \underbrace{\int_{\rho \in Q'} f_i(\alpha, \rho) \prod_{\substack{k=1 \\ k \neq i}}^L F_k(\rho) d\rho}, \quad (8.17)$$

where $f_i(\alpha, \rho)$ is the JPDF of $\alpha_{D,i}$ and ρ_i at the i th branch, $f_{\dot{\alpha}_{D,i}}(\dot{\alpha})$ is the PDF of $\dot{\alpha}_{D,i}$, the time derivative of $\alpha_{D,i}$, which is independent of $\alpha_{D,i}$ and $\alpha_{I,i}$ (therefore of ρ_i), and $Q' = \{\alpha_D = \alpha \cap 0 < \rho < \infty\}$.

3.2.2 Application to Different SC Algorithms.

Best Signal Power Algorithm. When the best signal power algorithm is employed at the SC combiner, the outage probability for dual-branch i.i.d. Rayleigh fading case can be shown by applying the method provided in last section to be given in closed-form as

$$P_{\text{out}}^S = [1 - \exp(-S_r)]^2 + 2 \sum_{n=0}^{N_I-1} \frac{\lambda_{rr}^n}{n!} \\ \times \left\{ \frac{\Gamma[n+1, S_r(1 + \lambda_{rr})]}{(1 + \lambda_{rr})^{n+1}} - \frac{\Gamma[n+1, S_r(2 + \lambda_{rr})]}{(2 + \lambda_{rr})^{n+1}} \right\}. \quad (8.18)$$

where $S_r = \frac{s_{\text{th}}}{\Omega_D}$ and $\lambda_{rr} = \frac{\Omega_D}{\lambda_{\text{th}} \Omega_I}$. Eqn. (8.18) can be shown to reduce to [9, Eqn. (5)] when $N_I = 1$, as expected. The LCR with a minimum power requirement $N(\lambda_{\text{th}}, s_{\text{th}})$ for the best signal power algorithm can

also be obtained in closed-form as

$$\begin{aligned} N_S(\lambda_{th}, s_{th}) = & 2\sqrt{2\pi}f_D\sqrt{S_r}[\exp(-S_r) - \exp(-2S_r)] \\ & \times \left[1 - \frac{\Gamma(N_I, S_r\lambda_{rr})}{\Gamma(N_I)} \right] + \frac{2\sqrt{2\pi}\lambda_{rr}^{N_I} \left(f_D^2 + \frac{f_L^2}{\lambda_{rr}} \right)^{\frac{1}{2}}}{\Gamma(N_I)} \\ & \times \left\{ \frac{\Gamma[N_I + \frac{1}{2}, S_r(1 + \lambda_{rr})]}{(1 + \lambda_{rr})^{N_I + \frac{1}{2}}} - \frac{\Gamma[N_I + \frac{1}{2}, S_r(2 + \lambda_{rr})]}{(2 + \lambda_{rr})^{N_I + \frac{1}{2}}} \right\}. \end{aligned} \quad (8.19)$$

Then the corresponding AOD $T_S(\lambda_{th}, s_{th})$ is just the ratio of (8.18) and (8.19) as per (8.2).

Special Cases

■ Interference-limited Case

The LCR $N(\lambda_{th})$ for interference-limited systems can be obtained by setting $s_{th} = 0$ in (8.19) which results into the compact formula

$$\begin{aligned} N_S(\lambda_{th}) = & \frac{2\sqrt{2\pi}\lambda_{rr}^{N_I}\Gamma(N_I + \frac{1}{2})\left(f_D^2 + \frac{f_L^2}{\lambda_{rr}}\right)^{\frac{1}{2}}}{\Gamma(N_I)} \\ & \times \left[\frac{1}{(1 + \lambda_{rr})^{N_I + \frac{1}{2}}} - \frac{1}{(2 + \lambda_{rr})^{N_I + \frac{1}{2}}} \right], \end{aligned} \quad (8.20)$$

and the corresponding P_{out} can be obtained by evaluating (8.18) at $s_{th} = 0$ as

$$P_{out}^S = 1 - 2 \left(\frac{\lambda_{rr}}{1 + \lambda_{rr}} \right)^{N_I} + \left(\frac{\lambda_{rr}}{2 + \lambda_{rr}} \right)^{N_I}. \quad (8.21)$$

Hence, as per (8.2), the resulting AOD is just the ratio of (8.21) and (8.20).

■ Power-limited Case

The outage probability of the power-limited case P_{out} can be obtained by evaluating (8.18) at $\lambda_{th} = 0$ or [36, Eqn. (6.7)] as

$$P_{out}^S = [1 - \exp(-S_r)]^2. \quad (8.22)$$

Setting $\lambda_{th} = 0$ in (8.19), we obtain the corresponding LCR $N(s_{th})$ of the signal power s_D at level s_{th} (purely power-limited case) which is given by

$$N_S(s_{th}) = 2\sqrt{2\pi}f_D\sqrt{S_r}[\exp(-S_r) - \exp(-2S_r)], \quad (8.23)$$

and the resulting AOD is

$$T_S(s_{\text{th}}) = \frac{\exp(S_r) - 1}{2\sqrt{2\pi}f_D\sqrt{S_r}}, \quad (8.24)$$

in agreement with [24, Eqns. (18a) and (20a)] for the dual-branch case (or equivalently [22, Eqns. (15) and (16)] and [23, Eqn. (18)] with $L = 2$).

Best SIR Algorithm. When the selection decision is based on the best SIR algorithm, the diversity branch with the highest SIR is selected. The best SIR algorithm usually provides the best performance for interference-limited systems, but is the most complex to implement. In this section, we study the performance of SC system based on the best SIR algorithm for both interference-limited and power-limited systems.

When the best SIR algorithm is employed at the SC combiner, the outage probability for dual-branch i.i.d. Rayleigh fading case can be shown to be given by

$$\begin{aligned} P_{\text{out}}^{\text{SIR}} = & 1 - 2 \exp(-S_r) \left[1 - \frac{\Gamma(N_I, S_r \lambda_{\text{rr}})}{\Gamma(N_I)} \right] - 2 \left(1 + \frac{1}{\lambda_{\text{rr}}} \right)^{-N_I} \frac{\Gamma[N_I, S_r(1 + \lambda_{\text{rr}})]}{\Gamma(N_I)} \\ & + 2N_I \sum_{k=0}^{N_I} \frac{1}{k!} \sum_{l=0}^{2N_I-1} \binom{2N_I-1}{l} (-1)^l S_r^{l+1} \\ & \times \{ \Gamma[k-l-1, S_r] - \Gamma[k-l-1, S_r(1 + \lambda_{\text{rr}})] \}. \end{aligned} \quad (8.25)$$

The corresponding LCR can be obtained as

$$\begin{aligned} N_{\text{SIR}}(\lambda_{\text{th}}, s_{\text{th}}) = & 2\sqrt{2\pi}f_D\sqrt{S_r} [\exp(-S_r) \\ & - \frac{1}{\Gamma(N_I)} \sum_{l=0}^{2N_I-1} \binom{2N_I-1}{l} (-S_r)^l \Gamma(N_I-l, S_r)] \\ & \times \left[1 - \frac{\Gamma(N_I, S_r \lambda_{\text{rr}})}{\Gamma(N_I)} \right] + \frac{2\sqrt{2\pi}\lambda_{\text{rr}}^{N_I} \left(f_D^2 + \frac{f_I^2}{\lambda_{\text{rr}}} \right)^{\frac{1}{2}}}{\Gamma(N_I)} \\ & \times \left\{ \frac{\Gamma[N_I + \frac{1}{2}, S_r(1 + \lambda_{\text{rr}})]}{(1 + \lambda_{\text{rr}})^{N_I + \frac{1}{2}}} - \frac{2}{\lambda_{\text{rr}}^{N_I + \frac{1}{2}} \Gamma(N_I)} \sum_{l=0}^{2N_I-1} \binom{2N_I-1}{l} \right. \\ & \left. \times \left(-\frac{1}{\lambda_{\text{rr}}} \right)^l \int_{\sqrt{s_{\text{th}}\lambda_{\text{rr}}}}^{\infty} y^{2l+2N_I} \exp(-y^2) \Gamma\left(N_I-l, \frac{y^2}{\lambda_{\text{rr}}}\right) dy \right\}. \end{aligned} \quad (8.26)$$

Note that the difference between the LCR results for best signal power algorithm and best SIR algorithm given in (8.19) and (8.26) is only in

the subtraction parts. Setting $s_{\text{th}} = 0/\lambda_{\text{th}} = 0$ in (8.25) and (8.26), one can get new results for the LCR and AOD for interference-limited case/power-limited case in closed-form [26].

Best S+I Algorithm. In the best S+I algorithm, the diversity branch with the highest total (desired plus interference) power is selected. This algorithm is the most practical to implement among the three algorithms. In the following subsection, the performance of SC system based on the best S+I algorithm for both interference-limited and power-limited systems will be investigated.

When the best SIR algorithm is employed at the SC combiner, the outage probability for independent not necessarily identically distributed dual-branch Rayleigh fading case can be shown to be given by [26]

$$P_{\text{out}}^{\text{SI}} = 1 - \sum_{\substack{i=1 \\ k=3-i}}^2 (I_1 - I_2 - I_3 + I_4), \quad (8.27)$$

where

$$\begin{aligned} I_1 &= \exp\left(\frac{s_{\text{th}}}{\Omega_{D,i}}\right) \left[1 - \frac{\Gamma\left(N_I, \frac{s_{\text{th}}}{\lambda_{\text{th}}\Omega_I}\right)}{\Gamma(N_I)} \right] \\ &\quad + \left(1 + \frac{\lambda_{\text{th}}\Omega_I}{\Omega_{D,i}}\right)^{-N_I} \frac{\Gamma\left(N_I, \frac{s_{\text{th}}}{\Omega_{D,i}} + \frac{s_{\text{th}}}{\lambda_{\text{th}}\Omega_I}\right)}{\Gamma(N_I)}, \end{aligned} \quad (8.28)$$

$$\begin{aligned} I_2 &= \frac{1}{2^{N_I}\Omega_{D,i}\Gamma(N_I)} \sum_{t=0}^{N_I-1} \frac{\left(\frac{1}{2}\right)^t}{t!} \sum_{z=0}^{N_I-1} \binom{N_I-1}{z} \left(-\frac{2}{\Omega_I}\right)^z \\ &\quad \times \Gamma(N_I+t-z) \sum_{m=0}^{N_I+t-z-1} \frac{\left(\frac{2}{\Omega_I}\right)^m}{m!} \left[\left(\frac{1}{\Omega_{D,i}} + \frac{1}{\Omega_I}\right)^{-m-z-1} \right. \\ &\quad \times \Gamma\left(m+z+1, \frac{s_{\text{th}}}{\Omega_{D,i}} + \frac{s_{\text{th}}}{\Omega_I}\right) - \left(\frac{1}{\Omega_{D,i}} + \frac{1}{\Omega_I} + \frac{2}{\Omega_I\lambda_{\text{th}}}\right)^{-m-z-1} \\ &\quad \times \left.\left(\frac{1+\lambda_{\text{th}}}{\lambda_{\text{th}}}\right)^m \Gamma\left(m+z+1, \frac{s_{\text{th}}}{\Omega_{D,i}} + \frac{s_{\text{th}}}{\Omega_I} + \frac{2s_{\text{th}}}{\lambda_{\text{th}}\Omega_I}\right)\right], \end{aligned} \quad (8.29)$$

$$\begin{aligned}
I_3 = & \frac{\left(1 - \frac{\Omega_I^2}{\Omega_{D,k}^2}\right)^{-N_I}}{\Omega_{D,i}\Gamma(N_I)} \sum_{z=0}^{N_I-1} \binom{N_I-1}{z} \left(-\frac{1}{\Omega_I} - \frac{1}{\Omega_{D,k}}\right)^z \\
& \times \Gamma(N_I-z) \sum_{m=0}^{N_I-z-1} \frac{\left(\frac{1}{\Omega_I} + \frac{1}{\Omega_{D,k}}\right)^m}{m!} \\
& \times \left[\left(\frac{1}{\Omega_{D,i}} + \frac{1}{\Omega_{D,k}}\right)^{-m-z-1} \Gamma\left(m+z+1, \frac{s_{\text{th}}}{\Omega_{D,i}} + \frac{s_{\text{th}}}{\Omega_{D,k}}\right) \right. \\
& - \left(\frac{1}{\Omega_{D,i}} + \frac{1+\lambda_{\text{th}}}{\Omega_{D,k}\lambda_{\text{th}}} + \frac{1}{\Omega_I\lambda_{\text{th}}}\right)^{-m-z-1} \left(\frac{1+\lambda_{\text{th}}}{\lambda_{\text{th}}}\right)^m \\
& \left. \times \Gamma\left(m+z+1, \frac{s_{\text{th}}}{\Omega_{D,i}} + \frac{s_{\text{th}}(1+\lambda_{\text{th}})}{\lambda_{\text{th}}\Omega_{D,k}} + \frac{s_{\text{th}}}{\lambda_{\text{th}}\Omega_I}\right) \right], \quad (8.30)
\end{aligned}$$

$$\begin{aligned}
I_4 = & \frac{\left(1 - \frac{\Omega_I}{\Omega_{D,k}}\right)^{-N_I}}{\Omega_{D,i}\Omega_I^{N_I}\Gamma(N_I)} \sum_{z=0}^{N_I-1} \frac{\left(\frac{1}{\Omega_I} - \frac{1}{\Omega_{D,k}}\right)^z}{z!} \sum_{t=0}^{N_I-1} \binom{N_I-1}{t} (-1)^t \\
& \times \left(\frac{2}{\Omega_I}\right)^{N_I+z-t} \Gamma(N_I+z-t) \sum_{m=0}^{N_I+z-t-1} \frac{\left(\frac{2}{\Omega_I}\right)^m}{m!} \left[\left(\frac{1}{\Omega_{D,i}} + \frac{1}{\Omega_I}\right)^{-m-t-1} \right. \\
& \times \Gamma\left(m+t+1, \frac{s_{\text{th}}}{\Omega_{D,i}} + \frac{s_{\text{th}}}{\Omega_I}\right) - \left(\frac{1}{\Omega_{D,i}} + \frac{1}{\Omega_I} + \frac{2}{\Omega_I\lambda_{\text{th}}}\right)^{-m-t-1} \\
& \left. \times \left(\frac{1+\lambda_{\text{th}}}{\lambda_{\text{th}}}\right)^m \Gamma\left(m+t+1, \frac{s_{\text{th}}}{\Omega_{D,i}} + \frac{s_{\text{th}}}{\Omega_I} + \frac{2s_{\text{th}}}{\lambda_{\text{th}}\Omega_I}\right) \right]. \quad (8.31)
\end{aligned}$$

The corresponding LCR for i.i.d Rayleigh case can be obtained as

$$\begin{aligned}
N_{\text{SI}}(\lambda_{\text{th}}, s_{\text{th}}) = & \sqrt{2\pi} f_D \sqrt{\Omega_D} (R_1 - R_2 - R_3 + R_4) |_{\alpha=\sqrt{s_{\text{th}}}} \\
& \times \left[1 - \frac{\Gamma(N_I, S_r \lambda_{\text{rr}})}{\Gamma(N_I)} \right] + \frac{4\sqrt{2\pi}}{\Gamma(N_I)\Omega_I^{N_I}} \sqrt{\frac{\lambda_{\text{th}}}{\Omega_D}} \left(f_D^2 + \frac{f_I^2 \Omega_I \lambda_{\text{th}}}{\Omega_D} \right)^{\frac{1}{2}} \\
& \times (A_1 - A_2 - A_3 + A_4), \quad (8.32)
\end{aligned}$$

where R_1-R_4 and A_1-A_4 are given by

$$R_1 = \frac{2\alpha}{\Omega_{D,i}} \exp\left(-\frac{\alpha^2}{\Omega_{D,i}}\right), \quad (8.33)$$

$$R_2 = \frac{\alpha}{2^{N_I-1}\Omega_{D,i}\Gamma(N_I)} \exp\left(-\frac{\alpha^2}{\Omega_{D,i}} + \frac{\alpha^2}{\Omega_I}\right) \sum_{t=0}^{N_I-1} \frac{\left(\frac{1}{2}\right)^t}{t!} \\ \times \sum_{z=0}^{N_I-1} \binom{N_I-1}{z} \left(-\frac{2\alpha^2}{\Omega_I}\right)^z \Gamma\left(N_I+t-z, \frac{2\alpha^2}{\Omega_I}\right), \quad (8.34)$$

$$R_3 = \frac{2\alpha \left(1 - \frac{\Omega_I^2}{\Omega_{D,k}^2}\right)^{-N_I}}{\Omega_{D,i}\Gamma(N_I)} \exp\left(-\frac{\alpha^2}{\Omega_{D,i}} + \frac{\alpha^2}{\Omega_I}\right) \sum_{z=0}^{N_I-1} \binom{N_I-1}{z} \\ \times \left(-\frac{\alpha^2}{\Omega_I} - \frac{\alpha^2}{\Omega_{D,k}}\right)^z \Gamma\left(N_I-z, \frac{\alpha^2}{\Omega_I} + \frac{\alpha^2}{\Omega_{D,k}}\right), \quad (8.35)$$

$$R_4 = \frac{\alpha \left(1 - \frac{\Omega_I}{\Omega_{D,k}}\right)^{-N_I}}{\Omega_{D,i} 2^{N_I-1} \Gamma(N_I)} \exp\left(-\frac{\alpha^2}{\Omega_{D,i}} + \frac{\alpha^2}{\Omega_I}\right) \sum_{z=0}^{N_I-1} \frac{\frac{1}{2^z} \left(1 - \frac{\Omega_I}{\Omega_{D,k}}\right)^z}{z!} \\ \times \sum_{t=0}^{N_I-1} \binom{N_I-1}{t} \left(-\frac{2\alpha^2}{\Omega_I}\right)^t \Gamma\left(N_I+z-t, \frac{2\alpha^2}{\Omega_I}\right), \quad (8.36)$$

$$A_1 = \frac{1}{2} \left(\frac{1}{\Omega_I} + \frac{\lambda_{\text{th}}}{\Omega_D} \right)^{-N_I - \frac{1}{2}} \Gamma\left(N_I + \frac{1}{2}, \frac{s_{\text{th}}}{\Omega_I \lambda_{\text{th}}} + \frac{s_{\text{th}}}{\Omega_D}\right), \quad (8.37)$$

$$A_2 = \frac{1}{2^{N_I+1}\Gamma(N_I)} \sum_{t=0}^{N_I-1} \frac{\left(\frac{1}{2}\right)^t}{t!} \sum_{z=0}^{N_I-1} \binom{N_I-1}{z} \left(-\frac{2\lambda_{\text{th}}}{\Omega_I}\right)^z \\ \times \Gamma(N_I+t-z) \sum_{m=0}^{N_I-t-z-1} \frac{\left(\frac{2\lambda_{\text{th}}}{\Omega_I}\right)^m}{m!} \left(\frac{\lambda_{\text{th}}}{\Omega_D} + \frac{\lambda_{\text{th}}}{\Omega_I} + \frac{1}{\Omega_I}\right)^{-N_I-m-z-\frac{1}{2}} \\ \times \Gamma\left(N_I+m+z+\frac{1}{2}, \frac{s_{\text{th}}}{\Omega_D} + \frac{s_{\text{th}}}{\lambda_{\text{th}} \Omega_I} + \frac{s_{\text{th}}}{\Omega_I}\right), \quad (8.38)$$

$$A_3 = \frac{\left(1 - \frac{\Omega_I^2}{\Omega_D^2}\right)^{-N_I}}{2\Gamma(N_I)} \sum_{z=0}^{N_I-1} \binom{N_I-1}{z} \left(-\frac{\lambda_{\text{th}}}{\Omega_I} - \frac{\lambda_{\text{th}}}{\Omega_D}\right)^z \\ \times \Gamma(N_I-z) \sum_{m=0}^{N_I-z-1} \frac{\left(\frac{\lambda_{\text{th}}}{\Omega_I} + \frac{\lambda_{\text{th}}}{\Omega_D}\right)^m}{m!} \left(\frac{1}{\Omega_I} + \frac{2\lambda_{\text{th}}}{\Omega_D}\right)^{-N_I-m-z-\frac{1}{2}} \\ \times \Gamma\left(N_I+m+z+\frac{1}{2}, \frac{2s_{\text{th}}}{\Omega_D} + \frac{s_{\text{th}}}{\lambda_{\text{th}} \Omega_I}\right), \quad (8.39)$$

$$\begin{aligned}
A_4 = & \frac{\left(1 - \frac{\Omega_I}{\Omega_D}\right)^{-N_I}}{2^{N_I+1}\Gamma(N_I)} \sum_{z=0}^{N_I-1} \frac{\frac{1}{2^z} \left(1 - \frac{\Omega_I}{\Omega_D}\right)^z}{z!} \sum_{t=0}^{N_I-1} \binom{N_I-1}{t} \left(\frac{2\lambda_{th}}{\Omega_I}\right)^t \\
& \times \Gamma(N_I+z-t) \sum_{m=0}^{N_I+z-t-1} \frac{\left(\frac{2\lambda_{th}}{\Omega_I}\right)^m}{m!} \left(\frac{\lambda_{th}}{\Omega_D} + \frac{\lambda_{th}}{\Omega_I} + \frac{1}{\Omega_I}\right)^{-N_I-m-t-\frac{1}{2}} \\
& \times \Gamma\left(N_I+m+t+\frac{1}{2}, \frac{s_{th}}{\Omega_D} + \frac{s_{th}}{\Omega_I} + \frac{s_{th}}{\lambda_{th}\Omega_I}\right). \tag{8.40}
\end{aligned}$$

3.2.3 Numerical Examples. Figures 8.6, 8.7, and 8.8 compare the system performance of dual-branch SC systems for all three algorithms with a minimum power requirement over i.i.d. Rayleigh diversity paths and in presence of multiple i.i.d. interfering Rayleigh faded signals. In particular, these figures show the comparisons of the outage probabilities, AODs, and average LCRs of SC as function of average SIR $\frac{\Omega_D}{\Omega_I}$ when $\frac{s_{th}}{\Omega_D} = -10$ dB, $\lambda_{th} = 8$ dB, and $N_I = 1$ and $N_I = 6$. They show that in the senses of outage probability and AOD, the best SIR algorithm provides the best performance over low average SIR range and the best signal power algorithm and best S+I algorithm, which have nearly the same performance for all three performance criteria over the entire range of average SIR shown in these figures, outperform the best SIR algorithm for high average SIR range. This is in agreement with the simulation results of outage probability reported in [9]. However, from Fig. 8.8, one can see that the best SIR algorithm almost always provides the worst performance for average LCR. Therefore in comparison to the best signal power algorithm and the best S+I, the best SIR algorithm yields error bursts with lower probability and shorter duration (in average). However the frequency of these bursts of errors is higher with the best SIR algorithm. It is also shown in these figures that the performance differences of SC for all three algorithms is greater when single interferer is present (i.e. $N_I = 1$).

3.3 Impact of Fading Correlation on the LCR and AOD of SC

Outage probability of SC over independent and correlated channels has been extensively studied over the past two decades (e.g., [4, 6, 18, 9] and references therein). On the other hand, the LCR and AOD of SC in the absence of CCI over various independent fading channels has been recently investigated in [24, 22, 23]. Pioneering work on the effect of correlated fading on the AOD and LCR of SC diversity systems was done by Adachi *et al.* [25]. More specifically, [25] presented a general method for

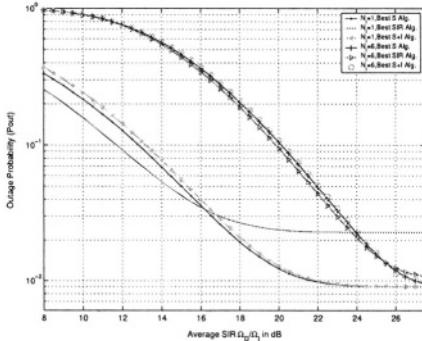


Figure 8.6. Comparison of the outage probabilities of SC for all three algorithms as function of average SIR $\frac{\Omega_D}{\Omega_I}$ when $\frac{s_{th}}{\Omega_D} = -10$ dB, $\lambda_{th} = 8$ dB, and $N_I = 1$ and $N_I = 6$.

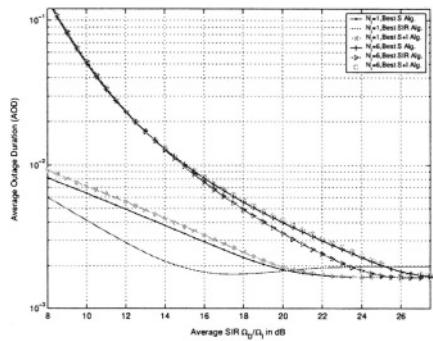


Figure 8.7. Comparison of the AODs of SC for all three algorithms as function of average SIR $\frac{\Omega_D}{\Omega_I}$ when $\frac{s_{th}}{\Omega_D} = -10$ dB, $\lambda_{th} = 8$ dB, and $N_I = 1$ and $N_I = 6$.

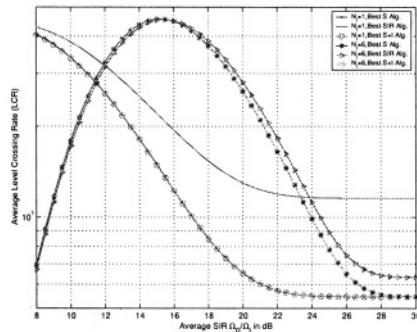


Figure 8.8. Comparison of the LCRs of SC for all three algorithms as function of average SIR $\frac{\Omega_D}{\Omega_I}$ when $\frac{s_{th}}{\Omega_D} = -10$ dB, $\lambda_{th} = 8$ dB, and $N_I = 1$ and $N_I = 6$.

calculating the LCR of correlated dual-branch SC system without CCI but ended up with approximate expressions for both the LCR and AOD. In [27], we complement these previous studies and present an exact and more general approach for the evaluation of the LCR and AOD of SC diversity systems with and without CCI operating over correlated and/or unbalanced multiple fading channels. The analysis is then specialized to a low-complexity dual-branch SC receiver over correlated Rayleigh and Nakagami fading channels. This particular scenario leads to relatively simple closed-form expressions and applies to small-size mobile termi-

nals equipped with space antenna diversity where the antenna spacing is insufficient to provide independent fading among the various signal paths.

3.3.1 General Method. We consider the SC systems employing the desired signal power algorithm. For communication systems which are both interference and power limited, the outage is defined in section 2.2. Outage probability of SC over correlated channels in presence of CCI with a minimal signal power requirement has been studied in [9] and closed-form expressions are available for the dual-branch i.i.d. diversity fading scenario combination of Rayleigh/Rayleigh in [9]. Therefore, to compute the AOD in (8.2), we just need to calculate the LCR $N(\lambda_{\text{th}}, s_{\text{th}})$ which can be obtained by using the method provided in section 2.2, except that the conditional JPDF $f_{r,\dot{r}|A}(r, \dot{r})$ for SC can be written as [26, Eqn. (20)]²

$$f_{r,\dot{r}|A}(r, \dot{r}) = \int_{\sqrt{\frac{s_{\text{th}}}{\lambda_{\text{th}}}}}^{\infty} \int_{-\infty}^{\infty} \alpha_I^2 f_{\alpha, \dot{\alpha}}(\alpha_I r, \dot{r} \alpha_I + \alpha_I \dot{r}) \\ \times f_{\alpha_I|A}(\alpha_I) f_{\dot{\alpha}_I}(\dot{\alpha}_I) d\alpha_I d\dot{\alpha}_I, \quad (8.41)$$

where the JPDF $f_{\alpha, \dot{\alpha}}(\alpha, \dot{\alpha})$ is given by

$$f_{\alpha, \dot{\alpha}}(\alpha, \dot{\alpha}) \\ = \sum_{l=1}^L f_{\dot{\alpha}_l}(\dot{\alpha}) \underbrace{\int_0^\alpha \int_0^\alpha}_{(L-1)-\text{fold}} f_{\alpha_1, \dots, \alpha_L}(\alpha_1, \dots, \alpha_l = \alpha, \dots, \alpha_L) \underbrace{d\alpha_1 \cdots d\alpha_k \cdots d\alpha_L}_{\substack{(L-1)-\text{fold} \\ k \neq l}} \\ \quad (8.42)$$

Furthermore, the LCR of purely power-limited SC system over correlated multiple channels can be obtained by applying Rice formula with (8.42) as

$$N(\alpha_{\text{th}}) = \sum_{l=1}^L \frac{\sigma_l}{\sqrt{2\pi}} \underbrace{\int_0^{\alpha_{\text{th}}} \int_0^{\alpha_{\text{th}}}}_{(L-1)-\text{fold}} f_{\alpha_1, \dots, \alpha_L}(\alpha_1, \dots, \alpha_l = \alpha_{\text{th}}, \dots, \alpha_L) \underbrace{d\alpha_1 \cdots d\alpha_k \cdots d\alpha_L}_{\substack{(L-1)-\text{fold} \\ k \neq l}} \\ = \sum_{l=1}^L \frac{\sigma_l}{\sqrt{2\pi}} \frac{\partial F_{\alpha_1, \dots, \alpha_L}(\alpha_1, \dots, \alpha_L)}{\partial \alpha_l} \Big|_{\{\alpha_k = \alpha_{\text{th}}\}_{k=1}^L}, \quad (8.43)$$

where $F_{\alpha_1, \dots, \alpha_L}(\alpha_1, \dots, \alpha_L)$ is the joint cumulative density function of $\alpha_1, \dots, \alpha_L$ and $\sigma_l^2 = \Omega_x \pi^2 f_m^2$, in which f_m is the maximum Doppler frequency shift, and $\Omega_x = \Omega_l$ for Rayleigh fading and $\Omega_x = \Omega_l/m$ for Nakagami fading.

3.3.2 Application to the Bivariate Nakagami Scenario.

Let the desired signal be subject to bivariate Nakagami- m fading with PDF provided by [17, Eqn.(6.1)]. It can be shown that the LCR with a minimum power requirement $N(\lambda_{\text{th}}, s_{\text{th}})$ is given by

$$\begin{aligned}
 & N(\lambda_{\text{th}}, s_{\text{th}}) \\
 = & \left\{ \sqrt{2\pi} f_m \frac{s_{\text{th}}^{m-0.5} m^{m-0.5}}{\Gamma(m) \Omega_1^{m-0.5}} \exp\left(-\frac{ms_{\text{th}}}{\Omega_1}\right) \left[1 - Q_m\left(\sqrt{\frac{2m\rho s_{\text{th}}}{(1-\rho)\Omega_1}}, \sqrt{\frac{2ms_{\text{th}}}{(1-\rho)\Omega_2}}\right) \right] \right. \\
 & + \sqrt{2\pi} f_m \frac{s_{\text{th}}^{m-0.5} m^{m-0.5}}{\Gamma(m) \Omega_2^{m-0.5}} \exp\left(-\frac{ms_{\text{th}}}{\Omega_2}\right) \left[1 - Q_m\left(\sqrt{\frac{2m\rho s_{\text{th}}}{(1-\rho)\Omega_2}}, \sqrt{\frac{2ms_{\text{th}}}{(1-\rho)\Omega_1}}\right) \right] \left. \right\} \\
 & \times \left[1 - \frac{\Gamma(N_I m_I, \frac{m_I s_{\text{th}}}{\Omega_I \lambda_{\text{th}}})}{\Gamma(N_I m_I)} \right] + \frac{\sqrt{2\pi}}{\Gamma(N_I m_I) \Gamma(m)} \left(f_m^2 + \frac{f_I^2 \lambda_{\text{th}} m \Omega_I}{m_I \Omega_1} \right)^{\frac{1}{2}} \\
 & \times \left\{ \frac{\Gamma(m+N_I m_I - \frac{1}{2}, \frac{m s_{\text{th}}}{\Omega_1} + \frac{m_I s_{\text{th}}}{\Omega_I \lambda_{\text{th}}})}{\left[1 + \left(\frac{m_I \Omega_1}{\lambda_{\text{th}} m \Omega_I} \right) \right]^{m+N_I m_I - \frac{1}{2}}} - 2 \left(\frac{m s_{\text{th}}}{\Omega_1} \right)^{m+N_I m_I - \frac{1}{2}} \right. \\
 & \times \int_1^\infty x^{2N_I m_I + 2m - 2} \exp\left[-\left(\frac{m s_{\text{th}}}{\Omega_1} + \frac{m_I s_{\text{th}}}{\Omega_I \lambda_{\text{th}}}\right)x^2\right] \\
 & \times Q_m\left(\sqrt{\frac{2m s_{\text{th}} \rho}{\Omega_1(1-\rho)}}x, \sqrt{\frac{2m s_{\text{th}}}{(1-\rho)\Omega_2}}x\right) dx \Bigg\} \\
 & + \frac{\sqrt{2\pi}}{\Gamma(N_I m_I) \Gamma(m)} \left(f_m^2 + \frac{f_I^2 \lambda_{\text{th}} m \Omega_I}{m_I \Omega_2} \right)^{\frac{1}{2}} \\
 & \times \left\{ \frac{\Gamma(m+N_I m_I - \frac{1}{2}, \frac{m s_{\text{th}}}{\Omega_2} + \frac{m_I s_{\text{th}}}{\Omega_I \lambda_{\text{th}}})}{\left[1 + \left(\frac{m_I \Omega_2}{\lambda_{\text{th}} m \Omega_I} \right) \right]^{m+N_I m_I - \frac{1}{2}}} - 2 \left(\frac{m s_{\text{th}}}{\Omega_2} \right)^{m+N_I m_I - \frac{1}{2}} \right. \\
 & \times \int_1^\infty x^{2N_I m_I + 2m - 2} \exp\left[-\left(\frac{m s_{\text{th}}}{\Omega_2} + \frac{m_I s_{\text{th}}}{\Omega_I \lambda_{\text{th}}}\right)x^2\right] \\
 & \times Q_m\left(\sqrt{\frac{2m s_{\text{th}} \rho}{\Omega_2(1-\rho)}}x, \sqrt{\frac{2m s_{\text{th}}}{(1-\rho)\Omega_1}}x\right) dx \Bigg\}, \tag{8.44}
 \end{aligned}$$

where $\rho = \frac{\text{cov}(\alpha_1^2, \alpha_2^2)}{\sqrt{\text{var}(\alpha_1^2) \text{var}(\alpha_2^2)}}$ is the correlation coefficient ($0 \leq \rho < 1$).

Special Cases

- Interference-limited Case

The LCR $N(\lambda_{\text{th}})$ for interference-limited systems can be obtained

by setting $s_{\text{th}} = 0$ in (8.44) and using an appropriate change of variable which results into

$$\begin{aligned}
 N(\lambda_{\text{th}}) = & \frac{\sqrt{2\pi} \left(\frac{m_I \Omega_1}{\lambda_{\text{th}} m \Omega_I} \right)^{N_I m_I}}{\Gamma(N_I m_I) \Gamma(m)} \left(f_m^2 + \frac{f_I^2 \lambda_{\text{th}} m \Omega_I}{m_I \Omega_1} \right)^{\frac{1}{2}} \left\{ \frac{\Gamma(m + N_I m_I - \frac{1}{2})}{\left[1 + \left(\frac{m_I \Omega_1}{\lambda_{\text{th}} m \Omega_I} \right) \right]^{m + N_I m_I - \frac{1}{2}}} \right. \\
 & - 2 \left(\frac{m \lambda_{\text{th}}}{\Omega_1} \right)^{m + N_I m_I - \frac{1}{2}} \int_0^\infty y^{2N_I m_I + 2m - 2} \exp \left[- \left(\frac{m \lambda_{\text{th}}}{\Omega_1} + \frac{m_I}{\Omega_I} \right) y^2 \right] \\
 & \times Q_m \left(\sqrt{\frac{2m \lambda_{\text{th}} \rho}{\Omega_1 (1-\rho)}} y, \sqrt{\frac{2m \lambda_{\text{th}}}{(1-\rho) \Omega_2}} y \right) dy \Big\} + \frac{\sqrt{2\pi} \left(\frac{m_I \Omega_2}{\lambda_{\text{th}} m \Omega_I} \right)^{N_I m_I}}{\Gamma(N_I m_I) \Gamma(m)} \\
 & \times \left(f_m^2 + \frac{f_I^2 \lambda_{\text{th}} m \Omega_I}{m_I \Omega_2} \right)^{\frac{1}{2}} \left\{ \frac{\Gamma(m + N_I m_I - \frac{1}{2})}{\left[1 + \left(\frac{m_I \Omega_2}{\lambda_{\text{th}} m \Omega_I} \right) \right]^{m + N_I m_I - \frac{1}{2}}} \right. \\
 & - 2 \left(\frac{m \lambda_{\text{th}}}{\Omega_2} \right)^{m + N_I m_I - \frac{1}{2}} \int_0^\infty y^{2N_I m_I + 2m - 2} \exp \left[- \left(\frac{m \lambda_{\text{th}}}{\Omega_2} + \frac{m_I}{\Omega_I} \right) y^2 \right] \\
 & \times Q_m \left(\sqrt{\frac{2m \lambda_{\text{th}} \rho}{\Omega_2 (1-\rho)}} y, \sqrt{\frac{2m \lambda_{\text{th}}}{(1-\rho) \Omega_1}} y \right) dy \Big\}. \tag{8.45}
 \end{aligned}$$

■ Power-limited Case

Setting $\lambda_{\text{th}} = 0$ in (8.44), we obtain the corresponding LCR $N(s_{\text{th}})$ of the signal power s_D at level s_{th} (purely power-limited case) which is given by the nicely compact and symmetrical closed-form expression

$$\begin{aligned}
 N(s_{\text{th}}) = & \sqrt{2\pi} f_m \frac{s_{\text{th}}^{\frac{2m-1}{2}} m^{m-0.5}}{\Gamma(m) \Omega_1^{m-0.5}} \exp \left(- \frac{ms_{\text{th}}}{\Omega_1} \right) \\
 & \times \left[1 - Q_m \left(\sqrt{\frac{2m \rho s_{\text{th}}}{(1-\rho) \Omega_1}}, \sqrt{\frac{2m s_{\text{th}}}{(1-\rho) \Omega_2}} \right) \right] \\
 & + \sqrt{2\pi} f_m \frac{s_{\text{th}}^{\frac{2m-1}{2}} m^{m-0.5}}{\Gamma(m) \Omega_2^{m-0.5}} \exp \left(- \frac{ms_{\text{th}}}{\Omega_2} \right) \\
 & \times \left[1 - Q_m \left(\sqrt{\frac{2m \rho s_{\text{th}}}{(1-\rho) \Omega_2}}, \sqrt{\frac{2m s_{\text{th}}}{(1-\rho) \Omega_1}} \right) \right]. \tag{8.46}
 \end{aligned}$$

The corresponding outage probability for dual-branch SC over correlated Nakagami fading is available in the form of a single integral

with finite limits in [17, Eqn. (6.26)]. As such taking a ratio of [17, Eqn. (6.26)] and (8.46) leads to the desired AOD result. Further setting $m = 1$ in (8.46) results in the LCR $N(\alpha_{\text{th}})$ for dual-branch SC over correlated Rayleigh fading.

- Independent and Balanced Rayleigh Channels Case

As a double check, setting $m = m_I = 1$, $\Omega_1 = \Omega_2 = \Omega_D$, and $\rho = 0$ in (8.44), and with the help of the identity $Q_1(0, b) = \exp(-b^2/2)$ given in [39, Eqn. (2)] and the definition of the incomplete gamma function, we obtain the LCR $N_{\text{Ray}}(\lambda_{\text{th}}, s_{\text{th}})$ of dual-branch SC over independent and balanced Rayleigh branches as

$$\begin{aligned} & N_{\text{Ray}}(\lambda_{\text{th}}, s_{\text{th}}) \\ &= 2\sqrt{2\pi}f_m\sqrt{S_r}[\exp(-S_r) - \exp(-2S_r)] \\ &\quad \times \left[1 - \frac{\Gamma(N_I, S_r\lambda_{\text{rr}})}{\Gamma(N_I)}\right] + \frac{2\sqrt{2\pi}\lambda_{\text{rr}}^{N_I}\left(f_m^2 + \frac{f_I^2}{\lambda_{\text{rr}}} \right)^{\frac{1}{2}}}{\Gamma(N_I)} \\ &\quad \times \left\{ \frac{\Gamma[N_I + \frac{1}{2}, S_r(1 + \lambda_{\text{rr}})]}{(1 + \lambda_{\text{rr}})^{N_I + \frac{1}{2}}} - \frac{\Gamma[N_I + \frac{1}{2}, S_r(2 + \lambda_{\text{rr}})]}{(2 + \lambda_{\text{rr}})^{N_I + \frac{1}{2}}} \right\}, \end{aligned} \quad (8.47)$$

where $S_r = \frac{s_{\text{th}}}{\Omega_D}$, $\lambda_{\text{rr}} = \frac{\Omega_D}{\lambda_{\text{th}}\Omega_I}$. Eqn.(8.47) is in perfect agreement with [26, Eqn. (36)].

3.3.3 Numerical Examples. As numerical examples, Figs. 8.9 and 8.10 show the effects of average power unbalance and fading correlation on the AOD and LCR of power-limited SC system. We can see that, compared to a system with i.i.d. fading over the diversity paths, both the average power unbalance and the fading correlation induce a noneligible degradation on the AOD performance. Furthermore, the effect of the average fading power unbalance is more significant than that of fading correlation for the typical parameters used in these figures. Finally, the impact of fading correlation on the AOD is more important when the branches tend to be balanced.

Figs. 8.11 and 8.12 plot the AOD and LCR of power-limited SC system versus normalized outage threshold of first path $\alpha_{\text{th}}/\sqrt{\Omega_1/m}$ over correlated Nakagami channels for various values of the correlation coefficient and of the desired user Nakagami parameter when $\Omega_1 = \Omega_2$. One can see that, when the desired user Nakagami parameter m increases (i.e., the fading severity deceases), the AOD decreases significantly (i.e., better performance) and the LCR curve shifts to the right. Similar to the observation to Figs. 8.9 and 8.10, Figs. 8.11 and 8.12 also shown that the effect of the desired user Nakagami parameter is more significant

than that of fading correlation. On the other hand, the impact of fading correlation on the AOD essentially remains the same as the Nakagami parameter is varied.

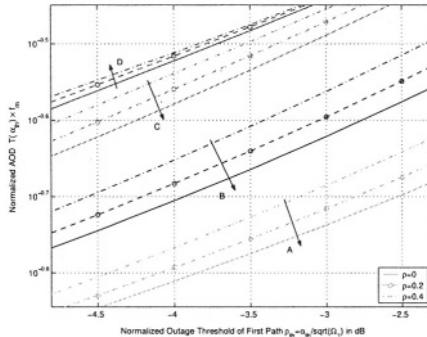


Figure 8.9. AOD of SC versus normalized outage threshold of first path $\alpha_{th}/\sqrt{\Omega_1}$ over correlated Rayleigh channels for various values of the correlation coefficient and of the average fading power unbalance. (A) $\Omega_1 = \Omega_2$, (B) $\Omega_1 = 2 \Omega_2$, (C) $\Omega_1 = 5 \Omega_2$, and (D) $\Omega_1 = 10 \Omega_2$.

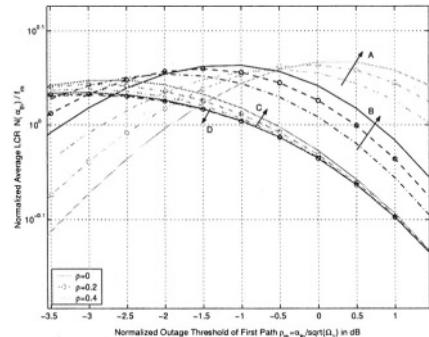


Figure 8.10. LCR of SC versus normalized outage threshold of first path $\alpha_{th}/\sqrt{\Omega_1}$ over correlated Rayleigh channels for various values of the correlation coefficient and of the average fading power unbalance. (A) $\Omega_1 = \Omega_2$, (B) $\Omega_1 = 2 \Omega_2$, (C) $\Omega_1 = 5 \Omega_2$, and (D) $\Omega_1 = 10 \Omega_2$.

Figs. 8.13 and 8.14 show the effects of the number of interferers and fading correlation on the AOD and LCR of SC. Specifically, they plot the AOD and LCR as a function of normalized SIR threshold of first path $\frac{m_I \Omega_1}{\lambda_{th} \Omega_1 m}$ over correlated Nakagami channels for various values of the correlation coefficient and of the number of interferers when $\Omega_1 = \Omega_2$, $m = m_I = 2$, $f_m = f_I = 40\text{Hz}$, and normalized desired signal power threshold of first path $\frac{s_{th} m}{\Omega_1} = -10 \text{ dB}$. Similarly we can see that as the number of interferers N_I increases, LCR curve shifts to the right again and AOD performance degrades considerably. Furthermore, for the chosen parameters the effect of the number of interferers is also more significant than that of fading correlation.

As the last set of examples, Figs. 8.15, 8.16, and 8.17 show the effects of minimum signal power constraint on the outage probability, AOD, and LCR of SC. From these figures, we can see that the minimum desired signal power requirement causes a floor on the outage probability, AOD, and LCR, since for $s_{th} > 0$ increasing the normalized SIR threshold above a particular value does not reduce the outage probability, AOD, and LCR, and the system becomes noise-limited. This floor and the

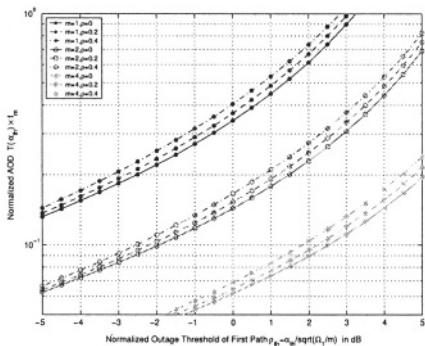


Figure 8.11. AOD of SC versus $\alpha_{th}/\sqrt{\Omega_1/m}$ over correlated Nakagami channels for various values of the correlation coefficient and of the desired user Nakagami parameter when $\Omega_1 = \Omega_2$.

value of the average SIR at which this floor occurs depends on $s_{th}m/\Omega_1$ for the fixed fading parameters.

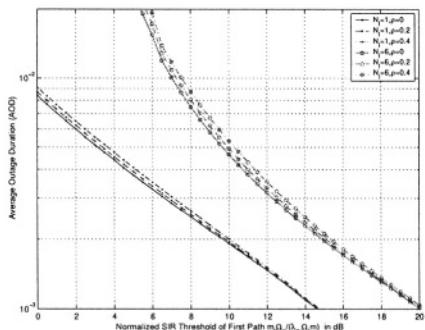


Figure 8.13. AOD of SC versus normalized SIR threshold of first path $\frac{m_0\Omega_1}{\lambda_{th}\Omega_1 m}$ over correlated Nakagami channels for various values of the correlation coefficient and of the number of interferers when $\Omega_1 = \Omega_2$, $m = m_I = 2$, $f_m = f_I = 40\text{Hz}$, and normalized desired signal power threshold of first path $\frac{s_{th}m}{\Omega_1} = -10\text{ dB}$.

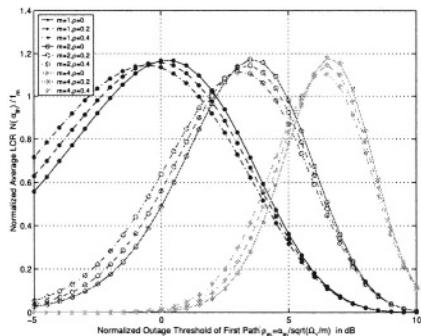


Figure 8.12. LCR of SC versus $\alpha_{th}/\sqrt{\Omega_1/m}$ over correlated Nakagami channels for various values of the correlation coefficient and of the desired user Nakagami parameter when $\Omega_1 = \Omega_2$.

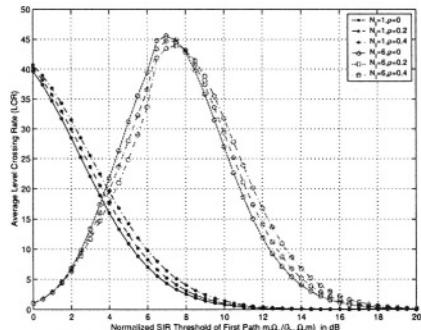


Figure 8.14. LCR of SC versus normalized SIR threshold of first path $\frac{m_0\Omega_1}{\lambda_{th}\Omega_1 m}$ over correlated Nakagami channels for various values of the correlation coefficient and of the number of interferers when $\Omega_1 = \Omega_2$, $m = m_I = 2$, $f_m = f_I = 40\text{Hz}$, and normalized desired signal power threshold of first path $\frac{s_{th}m}{\Omega_1} = -10\text{ dB}$.

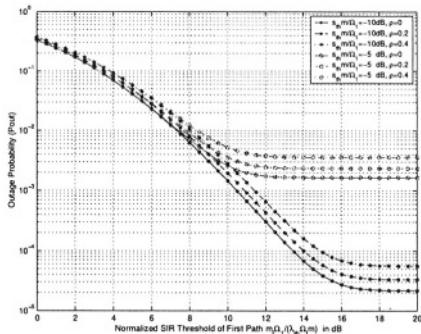


Figure 8.15. Outage Probability of SC $\frac{m_I\Omega_1}{\lambda_{th}\Omega_I m}$ over correlated Nakagami channels for various values of the correlation coefficient and of $\frac{s_{th}m}{\Omega_1}$ when $\Omega_1 = \Omega_2$, $m = m_I = 2$, $f_m = f_I = 40\text{Hz}$, and $N_I = 1$.

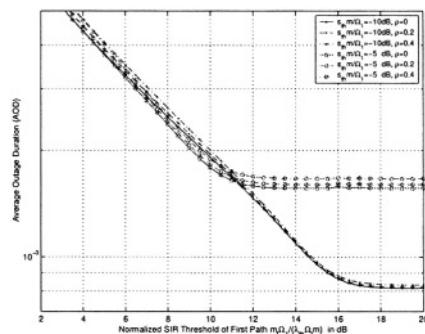


Figure 8.16. AOD of SC versus $\frac{m_I\Omega_1}{\lambda_{th}\Omega_I m}$ over correlated Nakagami channels for various values of the correlation coefficient and of $\frac{s_{th}m}{\Omega_1}$ when $\Omega_1 = \Omega_2$, $m = m_I = 2$, $f_m = f_I = 40\text{Hz}$, and $N_I = 1$.

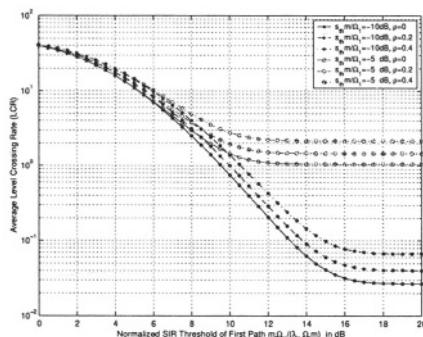


Figure 8.17. LCR of SC versus $\frac{m_I\Omega_1}{\lambda_{th}\Omega_I m}$ over correlated Nakagami channels for various values of the correlation coefficient and of $\frac{s_{th}m}{\Omega_1}$ when $\Omega_1 = \Omega_2$, $m = m_I = 2$, $f_m = f_I = 40\text{Hz}$, and $N_I = 1$.

4. Conclusion

In this chapter, we presented closed-form expressions for the LCR and AOD of some diversity combining schemes of interest with and without multiple CCI and over independent, correlated, and/or unbalanced channels considering minimum SIR and/or desired signal power constraints. More specifically, it presented generic results for the LCR and AOD

(i) of MRC systems subject to CCI operating over i.i.d. Rician and/or Nakagami fading environments when a minimum desired signal power requirement is specified for satisfactory reception, (ii) of various SC diversity scheme in presence of multiple CCI and with both minimum SIR and desired signal power constraints over independent, correlated, and/or unbalanced channels. Our numerical examples showed that: (i) Specifying a certain minimum desired signal power requirement induces a floor on the AOD, (ii) The AOD is primarily affected by the maximum Doppler frequencies (or equivalently the speed) and the amount of fading of the desired users, (iii) In the senses of outage probability and average outage duration, the best SIR algorithm provides the best performance in the low average SIR range while the best signal power algorithm and best S+I algorithm, which have nearly the same performance for all three performance criteria over the entire range of average SIR, outperform the best SIR algorithm for high average SIR range, and (iv) For typical parameters of interest the fading correlation has a nonnegligible impact on the system performance but its impact is less significant than that of the average fading power unbalance, desired user severity of fading, or the number of interferers.

Notes

1. By Nakagami/Rician, we mean that the desired user is subject to Nakagami fading, while the interferers are subject to Rician fading.
2. For simplicity, we use in this section α_i to stand for $\alpha_{D,i}$, Ω_i to represent $\Omega_{D,i}$, and α to be α_D .

References

- [1] K. W. Sowerby and A. G. Williamson, "Outage probability calculations for multiple cochannel interferers in cellular mobile radio systems," *IEE Proc. (Pt. F)*, vol. 135, no. 3, pp. 208–215, June 1988.
- [2] ———, "Selection diversity in multiple interferer mobile radio systems," *IEE Electron. Lett.*, vol. 24, pp. 1511–1513, November 1988.
- [3] A. A. Abu-Dayya and N. C. Beaulieu, "Outage probabilities of cellular mobile radio systems with multiple Nakagami interferers," *IEEE Trans. Veh. Tech.*, vol. VT-40, no. 4, pp. 757–768, November 1991.
- [4] S. Okui, "Probability of co-channel interference for selection diversity reception in the Nakagami m -fading channel," *IEE Proc I*, vol. 139, no. 1, pp. 91–94, February 1992.
- [5] Y. -D. Yao and A. U. H. Sheikh, "Investigation into cochannel interference in microcellular mobile radio systems," *IEEE Trans. Veh. Tech.*, vol. VT-41, no. 2, pp. 114–123, May 1992.

- [6] A. A. Abu-Dayya and N. C. Beaulieu, "Outage probabilities of diversity cellular systems with cochannel interference in Nakagami fading," *IEEE Trans. Veh. Tech.*, vol. VT-41, no. 4, pp. 343–355, November 1992.
- [7] H. -C. Yang and M. -S. Alouini, "Outage probability of wireless communication systems with a minimum signal power constraint: Some closed-form expressions," *IEEE Trans. Veh. Tech.*, vol. 51, pp. 1689–1698, Nov. 2002.
- [8] S. Okui, "Effects of CIR selection diversity with two correlated branches in the m -fading channel," *IEEE Trans. on Commun.*, vol. COM-48, pp. 1631–1633, October 2000.
- [9] H. -C. Yang and M. -S. Alouini, "Outage probability of dual-branch diversity systems in presence of co-channel interference," in *Proc. of the 12th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC'2001), San Diego, CA*, Sept. 2001, pp. 113–117, full journal paper in *IEEE Trans. Wireless Comm.*, vol. 2, pp. 310–319, Mar. 2003.
- [10] J. -P. M. G. Linnartz and R. Prasad, "Threshold crossing rate and average non-fade duration in a Rayleigh-fading channel with multiple interferers," *Archiv Fur Elektronik und Ubertragungstechnik Electronics and Communication*, vol. 43, pp. 345–349, Nov./Dec. 1989.
- [11] A. Abrardo and D. Sennati, "Outage statistics in CDMA mobile radio systems," *IEEE Commun. Letters*, vol. 4, no. 3, pp. 83–85, Mar. 2000.
- [12] Z. Cao and Y. -D. Yao, "Definition and derivation of level crossing rate and average fade duration in an interference-limited environment," in *Proc. IEEE Veh. Tech. Conf. (VTC'Fall 01), Atlantic City, NJ*, Oct. 2001, pp. 1608–1611.
- [13] L. Yang and M. -S. Alouini, "Average outage duration of multiuser wireless communication systems with a minimum signal power requirement," in *Proc. IEEE 55th Veh. Tech. Conf. (VTC'Spring 02), Birmingham, AL*, May 2002, pp. 1507–1511, full journal paper to appear in *IEEE Trans. Wireless Comm.*
- [14] F. Graziosi and F. Santucci, "Distribution of outage intervals in macrodiversity cellular systems," *IEEE Jour. Sel. Areas Commun.*, vol. 17, no. 11, pp. 2011–2021, Nov. 1999.
- [15] D. Brennan, "Linear diversity combining techniques," *Proc. IRE*, vol. 47, no. 6, pp. 1075–1102, June 1959.

- [16] Y. Roy, J. -Y. Chouinard, and S. A. Mahmoud, "Selection diversity combining with multiple antennas for mm-wave indoor wireless channels," *IEEE J. Select. Areas Commun.*, vol. 14, no. 4, pp. 674–682, May 1996.
- [17] M. K. Simon and M. -S. Alouini, *Digital Communications over Generalized Fading Channels: A Unified Approach to Performance Analysis*. Wiley, New York, 2000.
- [18] ———, "A unified performance analysis of digital communications with dual selective combining diversity over correlated Rayleigh and Nakagami- m fading channels," *IEEE Trans. Commun.*, vol. COM-47, no. 1, pp. 33–43, January 1999, see also *Proc. Communication Theory Mini-Conference (CTMC-VII) in conjunction with IEEE Global Commun. Conf. (GLOBECOM'98)*, pp. 28-33, Sydney, Australia, November, 1998.
- [19] M. -S. Alouini and M. K. Simon, "An MGF-based performance analysis of generalized selection combining over Rayleigh fading channels," *IEEE Trans. Commun.*, vol. 48, no. 3, pp. 401–415, March 2000.
- [20] Y. -C. Ko, A. Abdi, M. -S. Alouini, and M. Kaveh, "Average outage duration of diversity systems over generalized fading channels," *IEEE Trans. Veh. Tech.*, vol. 51, pp. 1672–1680, Nov. 2002.
- [21] X. Dong and N. C. Beaulieu, "Average level crossing rate and average fade duration of MRC and EGC diversity," in *Proc. IEEE Intern. Conf. Commun. (ICC'2001), Helsinki, Finland*, June 2001, pp. 1078–1083.
- [22] M. D. Yacoub, C. R. C. M. da Silva, and J. E. V. Bautista, "Second-order statistics for diversity-combining techniques in Nakagami-fading channels," *IEEE Trans. Veh. Tech.*, vol. 50, no. 6, pp. 1464–1470, Nov. 2001.
- [23] C. Iskander and P. Mathiopoulos, "Analytical level crossing rates and average fade durations for diversity techniques in Nakagami fading channels," *IEEE Trans. on Commun.*, vol. COM-50, no. 8, pp. 1301–1309, August 2002.
- [24] X. Dong and N. C. Beaulieu, "Average level crossing rate and average fade duration of selection diversity," *IEEE Commun. Letters*, vol. 5, no. 10, pp. 396–398, Oct. 2001.
- [25] F. Adachi, M. T. Feeney, and J. D. Parsons, "Effects of correlated fading on level crossing rates and average fade durations with pre-detection diversity reception," *IEE Proc. (Pt. F)*, vol. 135, no. 1, pp. 11–17, February 1988.

- [26] L. Yang and M. -S. Alouini, "Performance comparison of different selection combining algorithms in presence of co-channel interference," submitted to *IEEE Trans. Veh. Tech.*, Dec. 2002.
- [27] L. Yang and M. -S. Alouini, "An exact analysis of the impact of fading correlation on the average level crossing rate and average outage duration of selection combining," in *Proc. IEEE 57th Veh. Tech. Conf. (VTC Spring 2003), Jeju, Korea*, April 2003, pp. 241–245, full journal paper submitted to *IEEE Trans. Wireless Comm.*, Nov. 2003.
- [28] A. Shah and A. M. Haimovich, "Performance analysis of maximal ratio combining and comparision with optimum combining for mobile radio communications with cochannel interference," *IEEE Trans. Veh. Tech.*, vol. 49, no. 4, pp. 1454–1463, July 2000.
- [29] M. K. Simon and M. -S. Alouini, "On the difference of two chi-square variates with application to outage probability computation," *IEEE Trans. on Commun.*, vol. 49, pp. 1946–1954, Nov. 2001.
- [30] V.A. Aalo and J. Zhang, "Performance analysis of maximal ratio combining in the presence of multiple equal-power cochannel interferers in a Nakagami fading channel," *IEEE Trans. Veh. Tech.*, vol. 50, no. 2, pp. 479–503, March 2001.
- [31] Y. D. Yao and A. U. H. Sheikh, "Outage probability analysis for microcell mobile radio systems with co-channel interferers in Rician/Rayleigh fading environment," *IEE Electron. Lett.*, vol. 26, pp. 864–866, June 1990.
- [32] J. R. Haug and D. R. Ucci, "Outage probability for microcellular radio systems in a Rayleigh/Rician fading environment," in *Proc. IEEE Int. Conf. on Commun. (ICC'92)*, June 1992, pp. 312.4.1–312.4.5.
- [33] J. -C. Lin, W. -C. Kao, Y. T. Su, and T. -H. Lee, "Outage and coverage considerations for microcellular mobile radio systems in a shadowed-Rician/shadowed-Nakagami environment," *IEEE Trans. Veh. Tech.*, vol. VT-48, no. 1, pp. 66–75, January 1999.
- [34] T. T. Tjhung and C. C. Chai, "Distribution of SIR and performance of DS-CDMA systems in lognormally shadowed Rician channels," *IEEE Trans. Veh. Tech.*, vol. VT-49, no. 4, pp. 1110–1125, July 2000.
- [35] M. D. Yacoub, J. E. V. Bautista, and L. G. R. Guedes, "On higher order statistics of the Nakagami- m distribution," *IEEE Trans. Veh. Tech.*, vol. 48, no. 3, pp. 790–794, May 1999.

- [36] G. L. Stüber, *Principles of Mobile Communication*, 2nd ed. Kluwer Academic Publishers, Norwell, Massachusetts, 2000.
- [37] G. W. Lank and L. S. Reed, “Average time to loss of lock for an automatic frequency control loop with two fading signals and a related probability density function,” *IEEE Trans. Inform. Theory*, vol. 12, no. 1, pp. 73–75, Jan. 1966.
- [38] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, corr. and enlarg. ed. Orlando, FL: Academic Press, 1980.
- [39] A. H. Nuttall, “Some integrals involving the Q function,” Research Report of the Naval Underwater Systems Center, 1972.
- [40] W. C. Jakes, *Microwave Mobile Communication*, 2nd ed. Piscataway, NJ: IEEE Press, 1994.
- [41] L. Schiff, “Statistical suppression of interference with multiple Nakagami fading channels,” *IEEE Trans. Veh. Tech.*, vol. VT-21, pp. 121–128, November 1972.

Chapter 9

ENHANCING TCP PERFORMANCE IN WIDE-AREA CELLULAR WIRELESS NETWORKS

Transport Level Approaches

Ekram Hossain and Nadim Parvez

University of Manitoba

Department of Electrical and Computer Engineering

Winnipeg, MB, Canada R3T 5V6

Tel: (204) 474 8908, Fax: (204) 261 4639

{ekram, parvez}@ee.umanitoba.ca

Abstract

Internet technology-based architectures and protocols for supporting multimedia traffic over wireless networks are evolving. Since TCP (Transmission Control Protocol)/IP (Internet Protocol) is the standard network protocol stack on the Internet, its use over the next-generation wireless mobile networks is a certainty. Performance of TCP would be one of the most critical issues in IP-based data networking over wireless links. This article presents a comprehensive study on the performances of the basic TCP variants (e.g., TCP Tahoe, TCP Reno, TCP New-Reno, SACK TCP, FACK TCP) in wide-area cellular wireless networks. For the basic TCP variants, an in-depth analysis of the transport-level system dynamics is presented based on computer simulations using *ns-2*. Impacts of variations in wireless channel error characteristics, number of concurrent TCP flows and wireless link bandwidth on the average TCP throughput and fairness performances are investigated. The maximum achievable throughput under window-based end-to-end transmission control is also evaluated and the throughput performances of TCP New-Reno, SACK TCP and FACK TCP are compared against this ideal TCP throughput performance. To this end, an overview of the major modifications to the basic TCP variants based on transport-level approaches to enhance TCP performance in wired-cum-wireless networks is presented.

Keywords:

Wired-cum-wireless networks, TCP, transport level approach, random and correlated error.

Introduction

Transport layer protocol such as TCP operates on an end-to-end basis and its performance is one of the most critical issues in data networking over wireless links. A transport layer protocol is responsible for managing end-to-end flow and congestion control, providing reliability, security and QoS (Quality of Service) [1]. In OSI model, application protocol data unit (APDU) is passed down to the transport layer which is then called the transport service data unit (TSDU). A transport protocol header is added to this TSDU to form a transport protocol data unit (TPDU) and then it is passed to the network layer as a network service data unit (NSDU)¹.

TCP is the transport layer protocol used along with the unreliable network layer protocol IP (Internet Protocol) in todays Internet for non-real-time applications. The use of TCP/IP as the network protocol stack in the future generation wireless networks will leverage the rapidly evolving Internet technology in the wireless domain and enable to provide seamless wide-area Internet service to mobile users [2].

Transport protocol segments are transmitted in a wireless network as radio link level frames over the air-interface. For example, let us consider a TCP connection in a wide-area cellular wireless network, such as a GPRS (General Packet Radio Service) network. In case of transmission to a mobile from a fixed host in the Internet, after the SNDCP (Subnetwork Dependent Convergence Protocol) module in the SGSN (Serving GPRS Support Node) receives a TCP segment, it passes it to the LLC (Link Level Control) protocol (Fig. 9.1) [3]. The LLC protocol segments it into radio frames for transmission over the wireless link. Therefore, the performance of TCP will depend on the performance and service provided by the underlying radio link layer. In this article, however, we do not explicitly consider the impact of radio link protocols, rather we emphasize on the end-to-end (or transport-level) approaches while evaluating the performance of TCP in a wide-area wireless environment.

TCP is a connection-oriented reliable transport protocol consisting of three phases of operations: connection setup, data transfer, and connection termination. The connection set up procedure uses a three-way handshake where the connection is established after both ends of the connection inform each other of the set up process. The connection termination procedure uses a four-way handshake where both the active and the passive terminators send their own set of two-way handshake

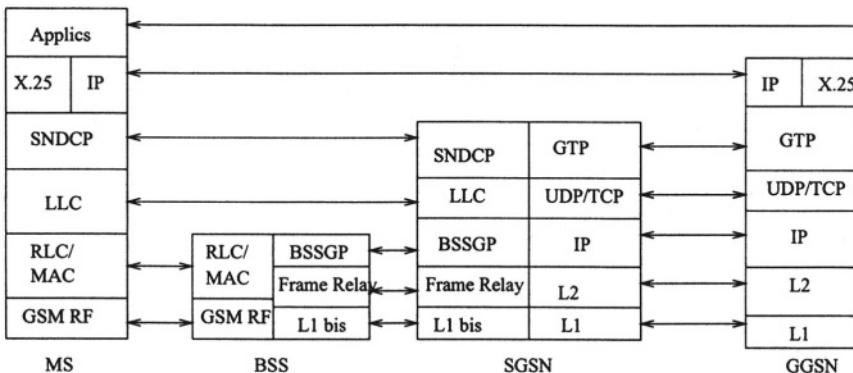


Figure 9.1. GPRS transmission protocol stack.

messages. Again, TCP is a bi-directional transport protocol that can transmit and receive at the same time (i.e., within the same connection set up). TCP is a byte-oriented transport protocol since it passes data to its peer in a byte by byte manner. For error control, TCP uses sequence number, cumulative positive acknowledgements with piggyback option for successfully transmitted packets and retransmission-based end-to-end error recovery. TCP uses a congestion control mechanism (to ensure network stability and prevent congestion collapse) which is intertwined with a window-based flow control mechanism. The TCP sender detects a packet-loss either by the arrival of several duplicate cumulative ACKs (Acknowledgements) or by the absence of an acknowledgement during a timeout interval and it attributes the packet loss to network congestion. Upon detection of a packet loss at the TCP sender, the congestion control/avoidance mechanism is triggered which reduces the transmission window size multiplicatively and/or increases the retransmission timer exponentially, and consequently, the throughput is reduced.

The implementation of the TCP flow/congestion control mechanism is based on two sender-side state variables, namely, *congestion window* ($cwnd$), $W(t)$ and the *slow-start threshold* ($ssthresh$), $W_{th}(t)$ ([4]-[6]). During connection set up, the receiver advertises a maximum window size W_{max} and the TCP sender is not allowed to have more than $\min(W_{max}, W(t))$ unacknowledged data packets outstanding at any given time t . For a new connection, $W(t)$ is generally initialized

to 1. The basic window adaptation mechanism (which is triggered by ACKs) in all currently available TCP implementations is as follows:

- i. If $W(t) < W_{th}(t)$ each ACK causes $W(t)$ to be incremented by 1. This is the *slow-start* phase.
- ii. If $W(t) \geq W_{th}(t)$ each ACK causes $W(t)$ to be incremented by $\frac{1}{(int)W(t)}$. This is the *congestion avoidance* phase.

In case of timeout, TCP source updates $W(t)$ and $W_{th}(t)$ as follows: $W_{th}(t+) = \lceil \frac{W(t)}{2} \rceil$ and $W(t+) = W_{th}(t+)$ and then starts retransmitting from the first lost packet. However, when the TCP source receives a cumulative acknowledgement from the TCP sink acknowledging the already transmitted packets, it immediately starts transmission after the highest ACKed packet. After a timeout the sender also updates the RTO (Retransmission Time Out) value using exponential backoff. The back-off continues until an ACK is received for a packet transmitted exactly once.

TCP sinks can accept packets out of order but deliver them only in sequence to the user and they generate immediate ACKs. If a sink receives a packet out of order, it issues an ACK immediately for the last in order packet that was received. If a packet is lost (after a stream of correctly received packets), then the receiver keeps sending ACKs (called duplicate ACKs) with the sequence number of the first lost packet even if packets transmitted after the lost packet are correctly received.

1. TCP in Wide-Area Cellular Wireless Environment

While the “traditional” TCP has been tuned for the last two decades to optimize its performance in wired networks where congestion is the main cause of any packet loss, it performs poorly in wireless environments due to the following factors:

- *Transmission losses on wireless link:* While in the wired transmission media the bit error rates are of $O(10^{-6}) - O(10^{-8})$, wireless media suffer significantly higher bit error rates such as $O(10^{-3}) - O(10^{-1})$ [7]. Therefore, the packet loss rate in a wireless link is an order of magnitude higher than that in a wired link. In a typical wireless scenario packet loss rate can be as high as 10% depending on channel fading and interference conditions and user mobility. When a packet is lost, TCP performs end-to-recovery by retransmitting the lost packet and frequent end-to-end retransmissions may reduce end-to-end throughput significantly. Again,

any packet loss in the wireless link is misinterpreted by the TCP sender as a congestion loss and it triggers the congestion control mechanism which reduces the sender's transmission window size resulting in reduced end-to-end throughput. Moreover, some of the mechanisms like *fast retransmit* proposed for enhancing the basic TCP algorithm may even fail in a wireless network due to the low bandwidth-delay product (and hence small TCP window size) and correlated channel fading (and hence loss of back-to-back packets and ACKs). Therefore, in a wired-cum-wireless environment, random packet losses in the wireless link cause the TCP sender to underestimate the available network bandwidth and thereby reduces the application layer performance.

Also, the location and time-dependent channel errors may impact the achieved throughput fairness among multiple concurrent TCP flows² and this would presumably be different for the different TCP variants. Again, frequent channel errors and subsequent retransmissions may result in inefficient use of the limited battery power in the mobile devices.

- *Wireless link delays:* Presence of limited bandwidth wireless links in an end-to-end path generally results in increased end-to-end transmission delays (and hence longer round-trip times (RTTs)). Although 2.5G and 3G cellular wireless networks (e.g., EGPRS, UMTS, IMT-2000) will have increased bandwidth, factors such as link asymmetry due to different uplink and downlink bandwidth, bandwidth oscillation due to dynamic resource allocation, stronger FEC (Forward Error Correction) coding, interleaving, radio link level recovery will result in large BDP (Bandwidth-Delay Product)³ wide-area cellular wireless networks [8].

Since the rate of increase in the congestion window size at the TCP sender is proportional to the rate of incoming ACKs, the congestion window may increase at a much lower rate in the presence of wireless links and thus reduce the end-to-end throughput. In fact, the throughput of a TCP connection has been shown to vary as the inverse of the connection's RTT. Given a packet loss rate p , the maximum sending rate for a TCP sender is γ bytes/sec, for $\gamma \leq \frac{1.5 * \sqrt{2/3} * B}{T * \sqrt{p}}$ where B is the packet size and T is the round trip time [9]. The inherent TCP bias against flows with longer RTT results in throughput unfairness among flows traversing the same number of hops but having different number of wireless links in the end-to-end path.

- *User mobility:* During a TCP session when a mobile user in a wireless cellular network moves from one cell to another, all necessary information must be transferred from the previous base station (BS) to the new base station which might cause a short duration of disconnection (typically of the order of several hundreds of milliseconds) during which no transmission takes place. Similar to that in the case of link errors, delays and packet losses during this “hand-off” scenario trigger TCP congestion control mechanisms which results in reduced end-to-end throughput [10].
- *Short-lived TCP flows:* Data services which are more transactional than streaming in nature (e.g., web browsing on a smart phone, email) usually involve transmission of a rather small amount of data, and therefore, tend to require short TCP connection duration. It is possible that the entire transmission is completed while the TCP sender is in the slow-start phase and thus resulting in the under utilization of the network capacity.

There has been a flurry of recent works ([10]-[38]) on improving TCP performance over wireless networks. A taxonomy of the different proposed mechanisms to improve TCP performance in wireless networks is shown in Fig. 9.2. This article provides a unified study of the different transport-layer approaches to TCP design for the wired-cum-wireless networks. The performances of the different TCP variants including the newer variants such as TCP SACK (Selective Acknowledgement), TCP Vegas, TCP FACK (Forward Acknowledgement) are evaluated in a wide-area cellular network in terms of TCP throughput, fairness (in the case of multiple competing TCP flows). We consider wireless link outage due to channel impairments only and we consider both the random and the correlated wireless channel error conditions.

To gain insight into the performance behaviors of the different TCP variants, system dynamics characterized by the *slow-start*, the *fast retransmit/fast recovery* and the *timeout* events is analyzed. Note that, while the fairness of TCP congestion avoidance among competing flows in a wide-area cellular wireless network can be maintained through improved traffic management techniques (e.g., in [9]) and/or modified congestion avoidance algorithms (e.g., in [39]) in the wired Internet, it may deteriorate in the wireless last hop due to time and location-dependent channel errors.

Also, we evaluate the maximum achievable throughput under a window-based end-to-end transmission control by simulation and compare the throughput performances of several basic TCP variants against this ideal TCP throughput performance. To provide a unified study, several approaches to wireless TCP design based on transport level modifications to the basic variants are also discussed in the related context.

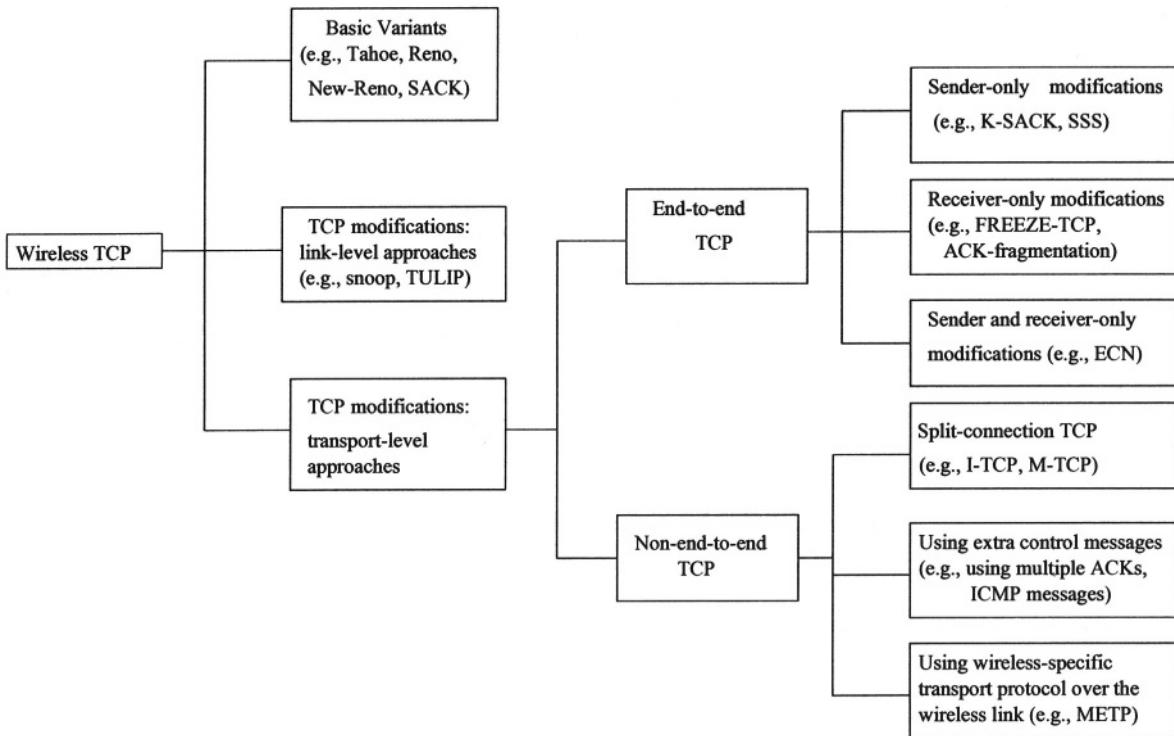


Figure 9.2. Taxonomy of TCP solutions (based on transport-level approaches) for wireless networks.

The motivation of this work is to investigate the effectiveness of different end-to-end TCP strategies in a wired-cum-wireless network under different wireless link conditions (e.g., random error, correlated error), and determine the limitations of TCP-like window-based flow control, which would be required to design more efficient end-to-end TCP. Note that, when IPSEC or other security mechanisms are employed to encrypt IP payloads, true end-to-end techniques based on transport level modifications of TCP would be required for wide-area wireless networks. In addition, it is feasible to implement end-to-end transport layer solutions in a wide-area wireless scenario where the connection end-point is a mobile host or a dedicated mobility/wireless-aware proxy since in this case the TCP implementations on stationary hosts need not be changed. Another concern is, when the data and the ACKs traverse different paths (e.g., in satellite networks), non-end-to-end solutions may cause serious problems, while end-to-end solutions are applicable over any type of networks and links.

For wide-area cellular wireless networks, end-to-end TCP solutions based on transport-level modifications which use some intermediary (such as a base station) may not maintain true end-to-end semantics (e.g., I-TCP [27]) and the intermediary may become a bottleneck due to the overhead involved in processing traffic associated with each connection and handing over the ‘state’ of each connection to the new intermediary during handoff.

Non-end-to-end TCP solutions such as those which use TCP-aware smarts in the base station (e.g., snoop protocol [13]) have scalability problems and may even degrade the performance of TCP in a wide-area cellular environment when the latency over the wireless link dominates the round trip time [22]. In addition, such a solution requires the base station to maintain significant state and is often tuned to specific flavor of TCP (e.g., snoop protocol does not work well with TCP Vegas).

Prior research closest to this work is that reported in [40], where the throughput performances of TCP Tahoe, OldTahoe, Reno and New-Reno in a localized wireless network were investigated under random packet losses using a stochastic model. The effects of coarse TCP timeout and the number of duplicate acknowledgements on TCP throughput performance were particularly emphasized. More general findings were reported in [41], where performance of TCP was investigated for a wide-area wireless network, particularly, for a network with high bandwidth-delay product and random packet losses. It was observed that random loss could lead to significant throughput degradation when the product of the packet loss (p) probability and the square of the bandwidth-delay product (i.e., $(\mu T)^2$, μ is the bandwidth and T is the delay) is larger than

one. Also, TCP's unfairness towards connections with higher round trip delays was reported. A packet level performance comparison among TCP Tahoe, Reno, New-Reno and SACK in a wide area wired-Internet environment was presented in [42] considering scenarios where the number of packet drops in a transmission window varies from one to four.

2. Basic TCP Variants

2.1 TCP Tahoe

TCP Tahoe uses *slow-start*, *congestion avoidance* and *fast retransmit* mechanisms [5]. During *slow-start*, the congestion window increases exponentially (by one for each acknowledgement received) until it reaches the slow-start threshold (*ssthresh*), and during *congestion avoidance* the congestion window increases linearly by one per round trip time (RTT). The TCP sender goes into the *fast retransmit* mode when it receives *tcp_rexmt_thresh* (which is usually set to 3) number of duplicate acknowledgements. During *fast retransmit*, the sender retransmits the lost segment and enters into the *slow-start* phase by setting the congestion window to 1 and *ssthresh* to the half of congestion window. In addition, it forgets all outstanding data transmitted earlier [42]. When the loss is due to sporadic channel error, switching to *slow-start* mode causes the throughput to fall.

2.2 TCP Reno

TCP Reno is similar to TCP Tahoe except that in addition to *fast retransmit*, it also includes the *fast recovery* mechanism for a single segment loss. When the TCP sender receives *tcp_rexmt_thresh* number of duplicate acknowledgements, instead of switching to *slow-start* after *fast retransmit*, TCP Reno enters into *fast recovery*. During *fast recovery*, the sender sets *ssthresh* to the half of the congestion window and the new congestion window to the new *ssthresh* plus the number of duplicate acknowledgements received. TCP Reno remains in *fast recovery* until the lost segment which triggered the *fast retransmit* has been acknowledged. When the sender receives new acknowledgement(s), it exits *fast recovery* and resets the congestion window to *ssthresh* and thereby moves into *congestion avoidance*.

In case of congestion loss, the *fast recovery* mechanism keeps the average congestion window size high resulting in better throughput performance compared to TCP Tahoe. During *fast recovery*, each new duplicate acknowledgement increases the congestion window size by one. Although TCP Reno works fine for single loss, in case of multiple losses

from the same transmission window the performance suffers since it exits *fast recovery* and enters into it again in a repeated fashion or goes to timeout.

2.3 TCP New-Reno

TCP New-Reno uses an augmented *fast recovery* mechanism where, unlike TCP Reno, *fast recovery* continues until all the segments which were outstanding during the start of the *fast recovery*, have been acknowledged [44]. This strategy helps to combat multiple losses without entering into *fast recovery* multiple times or causing timeout. In this case, a *partial acknowledgement*⁴ is considered as an indication that the segment following the acknowledged one has been dropped from the same transmission window (or flight), and therefore, TCP New-Reno immediately retransmits the other lost segment indicated by the *partial acknowledgement* and remains in *fast recovery*. It takes one round trip time to detect each lost segment and to retransmit it.

2.4 SACK TCP

In SACK TCP, the receiver sends acknowledgements with SACK (Selective Acknowledgement) option when it receives out of order segments due to loss or out of order delivery [45]. The SACK option field contains a number of SACK blocks, where each SACK block reports a non-contiguous set of data that has been received and queued. The first block in SACK option reports the most recently received block.

The SACK TCP sender is an intelligent extension of that in TCP Reno. It only modifies the *fast recovery* mechanism of TCP Reno keeping the other mechanisms unchanged [46]. Similar to New-Reno, it can handle multiple packet losses from the same flight. It has a better estimation capability for the number of outstanding segments. The acknowledgement with SACK option enables the sender to determine explicitly which segments have been received or have been lost. To keep track of the acknowledged and lost segments, it maintains a data structure called *Scoreboard*. Whenever the sender is allowed to transmit based on the congestion window size and the number of outstanding segments, it consults the *Scoreboard* and transmits the missing segments. If there is no missing segment to retransmit, it transmits new segments. When a retransmitted segment is dropped, the sender detects it by a retransmit timeout. In case of timeout it retransmits the segment and enters into the *slow-start* phase.

SACK TCP maintains a variable called *pipe* to keep track of the number of outstanding segments. For each retransmission or new transmis-

sion, the sender increases *pipe* by one and for each received acknowledgement, it decreases *pipe* by one. For each received partial acknowledgement, *pipe* is decreased by two with the assumption that the original packet and the retransmitted packet have left the network. Since *pipe* represents the amount of outstanding data, the sender transmits only when *pipe* is less than the congestion window.

2.5 FACK TCP

FACK (Forward Acknowledgement) TCP is a variant of SACK TCP with a modified *fast recovery* mechanism [46]. It uses a better technique for estimating the number of outstanding segments. For this, it introduces two new variables *snd.fack* and *retran-data*, where *snd.fack* represents the forward most data held by the receiver and *retran-data* represents the size of the retransmitted data outstanding in the network.

In non-recovery states *snd.fack* is updated using the acknowledgement number in the TCP header and is equivalent to *snd.una* [4]. But when a SACK block is received during error recovery, it is updated to the highest sequence number received by the receiver plus one regardless of the number of the intermediate dropped segments. Therefore, the outstanding data is estimated as $(\text{snd.nxt} - \text{snd.fack}) + \text{retran.data}$.

For each retransmission, the size of the *retran.data* is incremented by the corresponding segment size. When retransmitted segment leaves the network it is decreased by the same size. FACK TCP compares its estimate for outstanding data against the congestion window size and decides on the number of transmissions. A timeout is forced in the case of loss of a retransmitted segment on the assumption that the congestion is persistent. It enters into the *fast recovery* mode when the sender receives *tcp_rexmt_thresh* number of duplicate acknowledgements or $(\text{snd.fack} - \text{snd.una} > \text{tcp_rexmt_thresh} * \text{MSS})^5$. In the case that several segments are lost but fewer than *tcp_rexmt_thresh* number of duplicate acknowledgements have been received, the latter condition triggers the *fast recovery* phase sooner. Similar to New-Reno and SACK, FACK TCP terminates its recovery phase upon receiving acknowledgement for all the segments that were outstanding during transition to *fast recovery*.

2.6 TCP Vegas

TCP Vegas [47] comes with a proactive congestion control mechanism in which network congestion is predicted based on the estimated *expected throughput* and the *actual throughput*. Expected throughput is estimated as *WindowSize/BaseRTT*, where *BaseRTT* is the minimum of all the measured RTTs and *WindowSize* is the size of the current con-

gestion window. The actual throughput is calculated from the RTT for a tagged segment and the number of segments transmitted within that RTT. TCP Vegas compares the difference between the expected and the actual throughput against two thresholds α and β (where $\alpha < \beta$) and adjusts the transmission window accordingly. If the difference is smaller than α , the congestion window is increased linearly (by one per RTT) under the assumption that there is unutilized bandwidth available in the network. On the other hand, if the difference is larger than β , the congestion window is decreased, while if the difference lies between α and β the congestion window remains unchanged.

TCP Vegas has a fine grained timer expiry calculation mechanism to support early switching to *fast retransmit*. For this, the sender reads and records the system clock each time a segment is transmitted. When an acknowledgement arrives, it reads the clock again and calculates the fine grained RTT. TCP Vegas uses this fine-grained RTT estimate to calculate RTO. For each duplicate acknowledgement received, the sender checks the *Vegas expiry*⁶ and if Vegas expiry occurs, the sender switches to *fast retransmit*. Similar to the other TCP variants, it switches to *fast retransmit* when it receives *tcp_rexmt_thresh* number of duplicate acknowledgements. Also, the sender switches to *slow-start* whenever the usual timeout occurs.

2.7 Comparison Among the Basic TCP Variants

Although the TCP senders for all the basic TCP variants rely on using binary positive acknowledgement, similar RTO estimation and timeout mechanisms, they use different congestion avoidance mechanisms. The major differences among the basic TCP variants are shown in Table 9.1.

The main difference between TCP Tahoe and TCP Reno is that the latter uses the *fast recovery* mechanism while the former does not. Although TCP Reno can handle single drop in a flight efficiently, it cannot handle multiple packet drops in a single flight very well. Tahoe does not memorize outstanding data when it switches to slow-start, but Reno does. Both TCP New-Reno and SACK TCP can handle multiple losses in a single flight more efficiently.

SACK TCP is different from TCP New-Reno in that SACK TCP can retransmit selectively. While TCP New-Reno retransmits starting from the segment corresponding to the duplicate acknowledgement or partial acknowledgement, SACK TCP retransmits this segment along with other missing segments from its *scoreboard*. That is, the retransmission pool for a New-Reno sender can have at most one segment regardless of the number of packet drops. Therefore, it takes one RTT to recover

Table 9.1. Differences among the basic TCP variants.

| TCP Variant | Sender mechanism(s) |
|-------------|--|
| Tahoe | No <i>fast recovery</i> |
| Reno | Uses <i>fast recovery</i> to recover from single packet drops |
| New-Reno | Uses <i>fast recovery</i> to recover from multiple packet drops |
| SACK | Uses <i>fast recovery</i> to recover from multiple packet drops, better outstanding data estimation mechanism and larger retransmission pool |
| FACK | Similar to SACK, but uses a better outstanding data estimation mechanism and switches to <i>fast recovery</i> faster |
| Vegas | Uses proactive mechanism to control congestion window, switches to <i>fast retransmit</i> faster due to the fine-grained timer |

each packet loss. On the other hand, since the retransmission pool for a SACK sender can have many lost segments, the recovery process becomes faster. Also, due to the *pipe* variable, SACK TCP has a better estimation of the number of outstanding segments.

The main advantage of FACK TCP over SACK TCP is that, by using *retran_data* and *snd.fack* the former performs a better estimation of the number of outstanding segments than the latter. Generally, the *pipe* estimation for FACK TCP becomes smaller compared to that for SACK TCP which allows the former to transmit more segments. Also, using the information from SACK block, sometimes it switches to *fast recovery* before receiving *tcp_rexmt_thresh* number of duplicate acknowledgements which is helpful when the transmission window size is small or the number of duplicate acknowledgements is too few to trigger *fast retransmit*.

By using *expected throughput*, *actual throughput* and the threshold parameters α , β , TCP Vegas can have a better estimation of the available bandwidth compared to the other TCP variants. Due to its own fine-grained timer management, TCP Vegas can switch to *fast retransmit* earlier which contributes to its performance improvement. Also, since it reduces the congestion window to $3/4$ instead of $1/2$ during *fast retransmit* (when the segment that triggered *fast retransmit* has not been transmitted more than once), it helps to combat losses due to the sporadic wireless channel errors more efficiently.

3. Analysis of System Dynamics and Performance Evaluation

3.1 Simulation Model and Performance Measures

System dynamics under the different basic TCP variants, namely, TCP Tahoe, Reno, New-Reno, SACK, FACK and Vegas are investigated using *ns-2* [48] under varying wireless channel bandwidth, round trip delay, number of concurrent TCP connections and different wireless channel error characteristics in a wired-cum-wireless scenario where the mobile nodes act as TCP sinks. Also, two variants of New-Reno – one with *TCP-aware link level retransmission* (TLLR) and the other with *delayed acknowledgement* (DACK)-capable TCP sink, are considered.

It is assumed that the slowly varying shadow and path losses in a wireless link are perfectly compensated and that the channel quality variations due to multipath fading remain uncompensated. For a broad range of parameters, the packet error process in a mobile radio channel, where the multipath fading is considered to follow a Rayleigh distribution, can be modeled using a two-state Markov chain with transition probability matrix M_c [49]: $M_c = \begin{bmatrix} p & 1-p \\ 1-q & q \end{bmatrix}$ where p and $1-q$ are the probabilities that the j th packet transmission is successful, given that the $(j-1)$ th packet transmission was successful or unsuccessful, respectively.

Also, the fading envelope is assumed to not change significantly during transmission of a TCP packet (which is of duration, say, t_{tcp}). The channel variation for each of the mobiles is assumed to be independent and determined by the two parameters p and q which, in turn, are determined by $f_{dt_{tcp}}$ ⁷, the normalized Doppler bandwidth and P_E , the average packet error probability. P_E describes the channel quality in terms of *fading margin* F ⁸. Different values of P_E and $f_{dt_{tcp}}$ result in different degree of correlation in the fading process. When $f_{dt_{tcp}}$ is small, the fading process is highly correlated; on the other hand, for large values of $f_{dt_{tcp}}$, the fading process is almost independent. For a certain value of the average packet error rate P_E , the error burst-length increases as user mobility (and hence $f_{dt_{tcp}}$) decreases.

The network topology used in the simulation is shown in Fig. 9.3. The different performance measures are: *average per-flow throughput* (γ) (i.e., average amount of data successfully transmitted to a TCP sink per unit time), *energy consumption* (e) (average amount of energy dissipated in a TCP sink for successful reception of 1 MB of data) and *throughput fairness* (f). For n concurrent TCP flows, the fairness index

f is calculated as follows⁹:

$$f = \frac{(\sum_{i=1}^n \gamma_i)^2}{n \times \sum_{i=1}^n \gamma_i^2}. \quad (9.1)$$

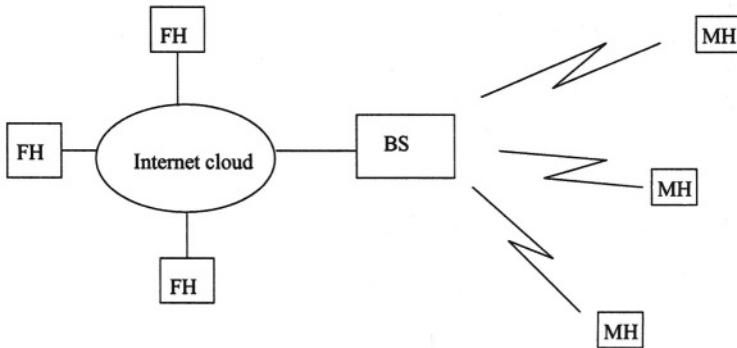


Figure 9.9. Simulation topology for wide-area cellular wireless network (FH ≡ Fixed host, MH ≡ mobile host, BS ≡ base station/wireless router).

To calculate the energy consumption at the mobile nodes (primarily due to transmission of acknowledgements and reception of TCP packets), the ratio of transmission power and receive power wattage is assumed to be 3 for all the simulation scenarios described in this article.

3.2 Simulation Results and Analysis

3.2.1 Scenario 1: Wireless link bandwidth = 2 Mbps, Internet delay = 20 ms, wireless link delay = 10 ms, single flow, maximum window size = 15, segment size = 1000 bytes, uniform wireless channel error. In this scenario, for the *delayed acknowledgement* case, the interval for delaying the acknowledgement is assumed to be 10 ms. The simulation run time is taken to be 1000 seconds.

Under random error conditions, TCP New-Reno and SACK TCP provide remarkable performance improvement over TCP Tahoe and TCP Reno (Fig. 9.4). Although TCP Tahoe and TCP Reno exhibit similar performance, TCP Reno performs slightly better than TCP Tahoe when the error rate is smaller than 1.5%. Analysis of the system events such as the total number of timeouts (TO), total number of *fast retransmit/fast recovery* (FR), number of *fast retransmits* due to multiple packet drop in a transmission window (MD), number of segments newly transmitted during *fast recovery* (SND), and average of the transmission

window sizes measured at the epoches of transitions to *fast retransmit* (FLT) reveal that under random error conditions the system dynamics for both TCP New-Reno and TCP SACK are more or less similar (Table 9.2), and therefore, their long-term performances are observed to be fairly close. FACK TCP is observed to perform better than each of TCP New-Reno and SACK TCP. TCP Vegas shows significant performance improvement over FACK TCP. However, the throughput performance of each of the above basic TCP variants is inferior to the performance of TCP New-Reno with TLLR (Fig. 9.4).

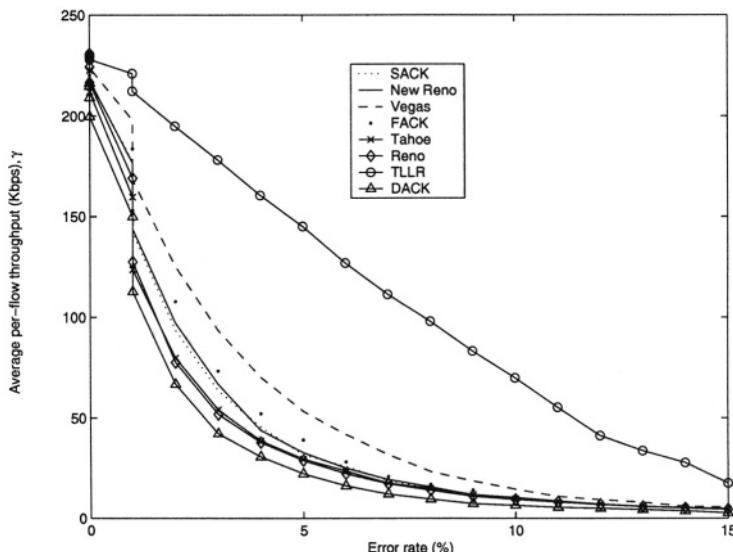


Figure 9.4. Throughput performances of different TCP variants (scenario 1).

For both TCP Tahoe and TCP Reno, the number of timeouts increases with increasing error rate, and the difference in the number of timeouts for these two cases during a certain observation interval is observed to be small. As the error rate increases, Reno suffers due to multiple packet drops in the same flight and also the number of timeouts increases compared to the number of *fast retransmits*. This offsets Reno's gain due to *fast recovery* and makes the performance closer to that of Tahoe.

As is observable from Table 9.2, at 1% error rate, for SACK TCP there are 115 events of *fast recovery* due to multiple packet drops (within a transmission window) among the total of 1252 *fast recovery events*, while for TCP New-Reno, among the total of 1237 *fast recovery events* 129 *fast recovery* events are due to multiple packet drops. Therefore, the ratio of *multiple drop fast recovery* to total *fast recovery* is smaller for SACK

Table 9.2. Fast retransmit/recovery analysis under random error loss (for 1000 second simulation run)

| TCP variant | TO | FR | MD | SND | FLT |
|------------------|-----|---------------|-----|------|-------|
| Error rate = 1% | | | | | |
| Tahoe | 53 | 1056 | | | |
| Reno | 100 | 1181 | 110 | 1909 | 11.59 |
| New-Reno | 27 | 1237 | 129 | 2013 | 11.92 |
| SACK | 25 | 1252 | 115 | 2435 | 11.88 |
| FACK | 17 | 1365 | 137 | 6668 | 12.27 |
| TLLR | 0 | 14 | 1 | 2 | 15.00 |
| Vegas | 22 | 1552 1072E | | | 13.04 |
| Error rate = 10% | | | | | |
| Tahoe | 691 | 295 | | | |
| Reno | 692 | 277 | 65 | 359 | 5.55 |
| New-Reno | 705 | 315 | 88 | 442 | 5.55 |
| SACK | 685 | 298 | 83 | 341 | 5.51 |
| FACK | 630 | 392 | 125 | 426 | 5.28 |
| TLLR | 20 | 95 | 58 | 134 | 13.94 |
| Vegas | 546 | 550 334E | | | 5.89 |

TCP. This observation holds for other error rates as well. In this case, SACK TCP cannot exploit the advantage of selective retransmission more effectively (compared to TCP New-Reno).

Again, under random error conditions, SACK TCP may not be able to exploit the advantage of better *pipe* estimation. Referring to Table 9.2, for 1% error rate SACK TCP transmits on the average 1.94 (= 2435/1252) new segments during *fast recovery*, while TCP New-Reno transmits on the average 1.63 (= 2013/1237) new segments. This small difference may not result in significant performance improvement for SACK TCP as compared to TCP New-Reno. As the error rate increases, the difference may become even smaller.

The *delayed acknowledgement* (DACK) variant of TCP New-Reno is observed to perform the worst, which is primarily due to its slow response to transmission failure.

Due to the better estimation of the number of outstanding segments, FACK TCP performs better than SACK TCP under all error conditions. At 1% and 10% error rate, for FACK TCP the number of new segments transmitted during *fast recovery* is observed to be 6668 and 426, respectively, while for TCP SACK, the numbers are 2435 and 341 (Table 9.2). With increasing error rate, as the effect of timeout becomes more dominant, the number of new segments transmitted during *fast recovery* decreases.

The performance improvement in TCP Vegas is mainly due to its two unique features – Vegas expiry and congestion window reduction by a factor of 3/4. The Vegas expiry mechanism causes it to react very fast to segment loss. As can be observed in Table 9.2, for 1% error rate, during 1000 second of simulation Vegas experiences 1552 *fast retransmits* among which 1072 were detected early by the Vegas fine-grained expiry mechanism (in Table 9.2, ‘E’ refers to fast retransmit detected by Vegas expiry). The *Vegas expiry* mechanism also reduces the number of timeouts and it is observed that among all the TCP variants it experiences the smallest number of timeouts.

The TCP New-Reno with TCP-aware link level retransmission (TLLR) scheme (e.g., snoop protocol [11]) offers the best performance (Fig. 9.4) due to the fact that the ‘snoop agent’ at the link level eliminates most of the fast retransmits and timeouts. For example, with TLLR, for 10% error rate, the number of fast retransmits and the number of timeouts are 20, and 95, respectively, while for New-Reno they are 705 and 315, respectively. As a result, the average flight size is always very high (e.g., 13.94 for 10% error rate) which is close to the assumed maximum transmission window size of 15, and consequently, better throughput performance is achieved.

Regarding energy consumption for transmission of a fixed amount of data (1 MB) to a mobile node using the different TCP variants, it is observed that TCP Vegas requires the highest amount of energy. This is due to the fact that, TCP Vegas results in the highest number of *fast retransmits*. For TCP New-Reno with the DACK option, the energy consumption is observed to be the lowest, which is due to its conservative acknowledgement transmission policy. But as the error rate increases, energy consumption due to reception of the retransmitted segments becomes more dominant compared to that due to transmission of acknowledgements, and consequently, energy consumption for TCP New-Reno with DACK tends to be similar to that for the other TCP variants (Fig. 9.5).

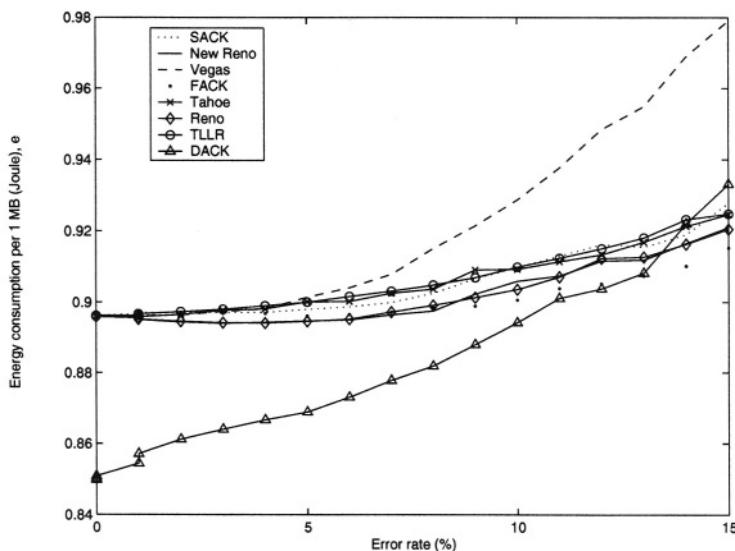


Figure 9.5. Energy consumption for the different TCP variants (scenario 1).

In fact, in the high error rate regime none of the end-to-end TCP mechanisms performs well in a wide-area cellular wireless environment. For example, with 10% error rate, the achieved thoroughput is about 10 Kbps compared to 217 Kbps in the ideal case. This is due to the inability of TCP to differentiate between wireless loss and congestion loss.

3.2.2 Scenario 2: Wireless link bandwidth = 10 Mbps, Internet delay = 20 ms, wireless link delay = 10 ms, 5 concurrent flows, maximum window size = 15, segment size = 1000 bytes,

uniform wireless channel error. System dynamics (in terms of timeout, fast retransmit/fast recovery) and energy consumption per flow under multiple concurrent TCP connections are similar to those in case of a single connection (Figs. 9.6-9.7). But the average per-flow throughput is observed to be slightly higher than the single flow case (scenario 1) when the error rate is not too high (e.g., $\leq 5\%$).

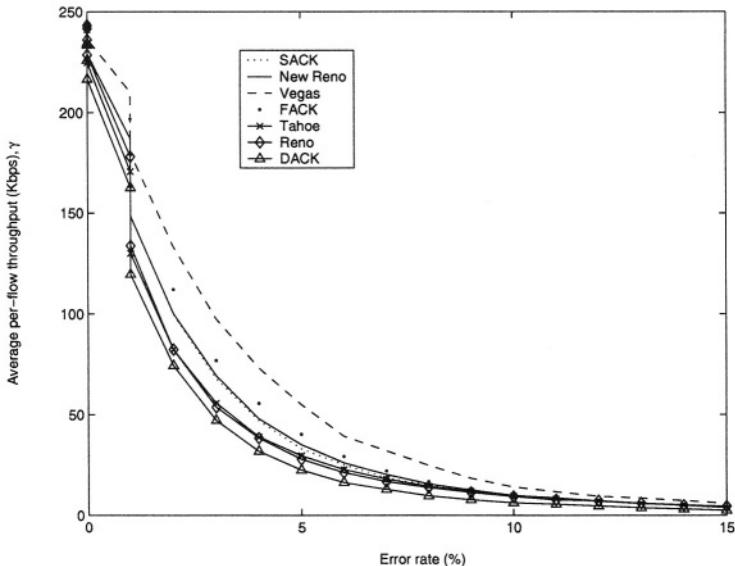


Figure 9.6. Throughput performances of different TCP variants (scenario 2).

In this case, we also observe the achieved throughput fairness among the different competing TCP connections. All of the TCP variants provide good long-term fairness and it is observed that even for error rate as large as 9%, the fairness index lies above 0.99. For error rate larger than 14%, the fairness index reduces to 98% (Fig. 9.8). Therefore, it can be concluded that the random wireless channel errors do not impact the TCP throughput fairness for the different TCP variants remarkably.

3.2.3 Scenario 3: Wireless link bandwidth = 10 Mbps, Internet delay = 20ms, wireless link delay = 10 ms, single flow, maximum window size = 75, segment size = 1000 bytes, uniform wireless channel error. This scenario is considered to compare the throughput performance in the case of multiple concurrent TCP flows with that in the case of a single TCP flow occupying the entire bandwidth. The average throughput of this case (which is

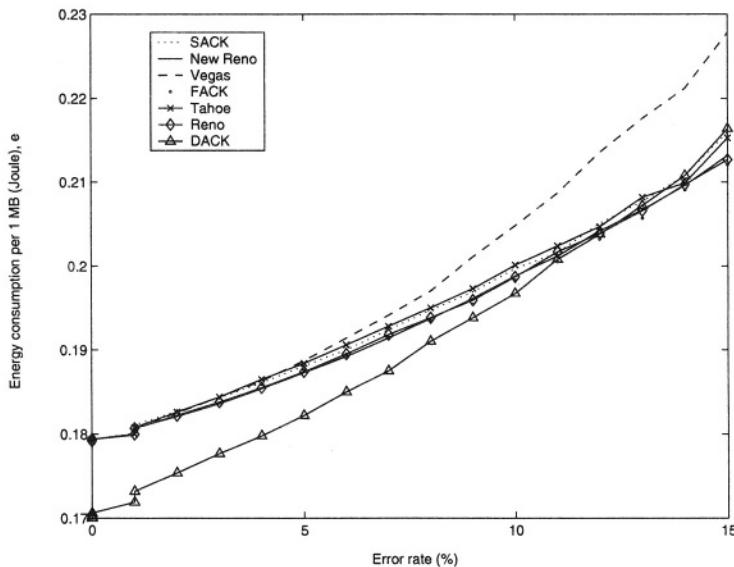


Figure 9.7. Energy consumption for the different TCP variants (scenario 2).

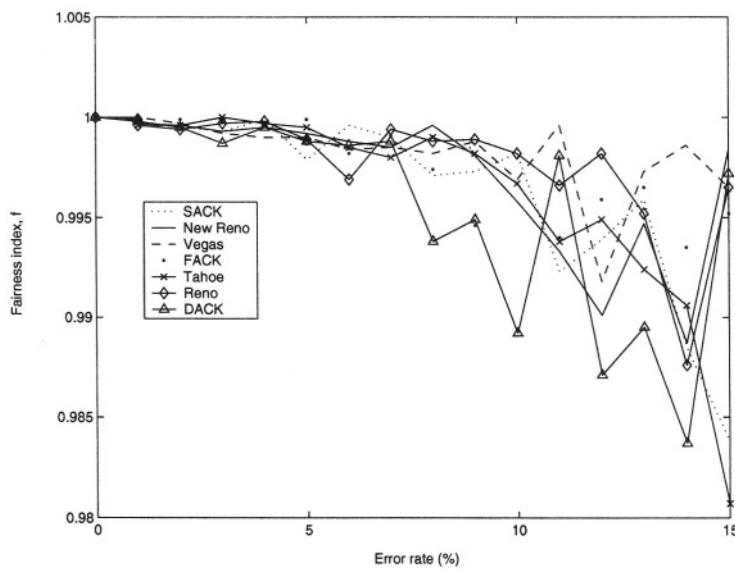


Figure 9.8. Throughput fairness for the different TCP variants (scenario 2).

expected to be five times that for a flow in scenario 2) becomes close to the per-flow throughput in scenario 2 as the error rate increases be-

yond 1% (Fig. 9.9). The performance gain due to the high wireless link bandwidth is observable only for very small error rates (e.g., 0%-0.05%).

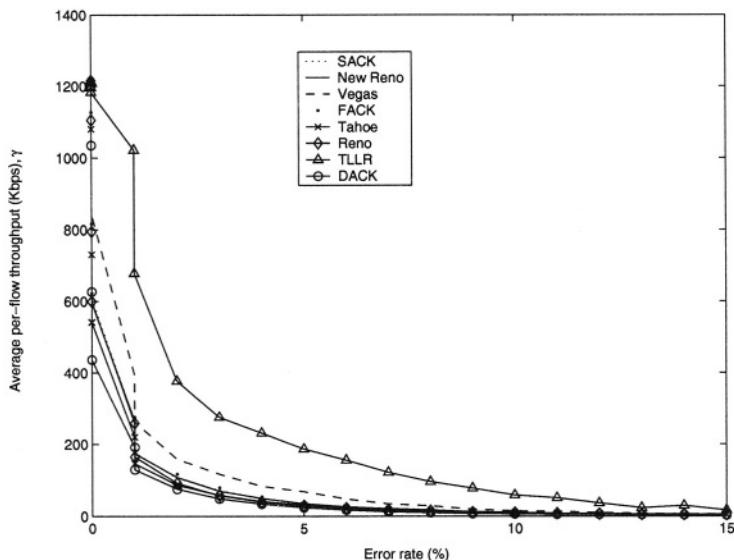


Figure 9.9. Throughput performances of different TCP variants (scenario 3).

3.2.4 Scenario 4: Wireless link bandwidth = 2 Mbps, Internet delay = 20 ms, wireless link delay = 10 ms, single flow, maximum window size = 15, segment size = 1000 bytes, correlated wireless channel error, TCP New-Reno only. In this scenario, we choose to investigate only the performance of TCP New-Reno (which is regarded as the de facto standard in the present Internet) under correlated channel error. For a particular average packet error rate, the channel error correlation (as manifested in the lengths of the error bursts) varies as the mobile speed varies [49]. For example, for the assumed wireless channel bandwidth and maximum segment size, with 10% error rate, mobile speed of 1 km/hr, 3 km/hr, 5 km/hr, 10 km/hr and 100 km/hr correspond to burst-error size of 11.41, 3.85, 2.37, 1.42 and 1.11 segments, respectively. For 1% average error rate, the corresponding burst-error sizes are 3.41, 1.39, 1.14, 1.04 and 1.01 segments, respectively.

The average throughput decreases with increased channel error correlation (Fig. 9.10). In the presence of highly correlated error, the acknowledgements from the mobile TCP sinks do not reach the TCP senders and the TCP error recovery is triggered primarily by timeouts. For example,

Table 9.3. Fast retransmit/fast recovery analysis under correlated error loss (for 1000 second run)

| Mobile speed | TO | FR | MD | SND | FLT |
|------------------|-----|-----|-----|-----|-------|
| Error rate = 3% | | | | | |
| 1 km/hr | 83 | 11 | 11 | 8 | 14.82 |
| 3 km/hr | 230 | 234 | 231 | 164 | 13.93 |
| 5 km/hr | 259 | 328 | 319 | 172 | 13.64 |
| 10 km/hr | 301 | 399 | 380 | 204 | 13.41 |
| 100 km/hr | 300 | 404 | 384 | 191 | 13.45 |
| Error rate = 5% | | | | | |
| 1 km/hr | 105 | 1 | 1 | 1 | 13.00 |
| 3 km/hr | 189 | 16 | 14 | 20 | 11.62 |
| 5 km/hr | 321 | 75 | 71 | 137 | 11.45 |
| 10 km/hr | 451 | 172 | 161 | 294 | 11.59 |
| 100 km/hr | 479 | 187 | 175 | 251 | 11.50 |
| Error rate = 10% | | | | | |
| 1 km/hr | 76 | 0 | 0 | | |
| 3 km/hr | 140 | 0 | 0 | | |
| 5 km/hr | 298 | 0 | 0 | | |
| 10 km/hr | 532 | 0 | 0 | | |
| 100 km/hr | 507 | 0 | 0 | | |

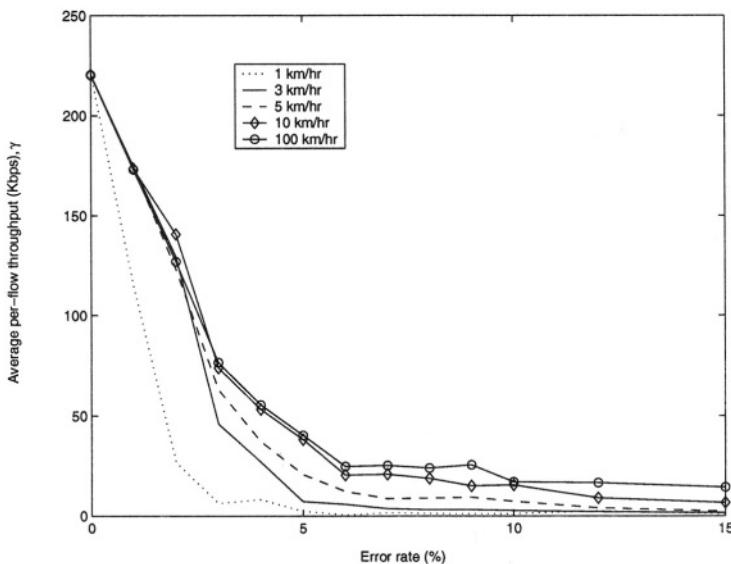


Figure 9.10. Throughput performance of TCP New-Reno (scenario 4).

with 10% error rate, it is observed that there is no *fast retransmit* at all (Table 9.3).

One interesting observation is that, although the number of timeouts decreases/increases with decreasing/increasing mobile speed, the average throughput decreases/increases. This is because, as user speed decreases, the channel error correlation increases, as a consequence of which the value of RTO generally becomes high. With an increase in the value of RTO, the end-to-end error recovery after a timeout becomes more sluggish, and consequently, the throughput deteriorates.

The amount of energy consumption is observed to increase with increasing channel error correlation.

3.2.5 Scenario 5: Wireless link bandwidth = 2 Mbps, Internet delay = 20 ms, wireless link delay = 10 ms, single flow, maximum window size = 15, segment size = 1000 bytes, correlated wireless channel error. We assume mobile speed of 3 km/hr at which the burst-error size is 3.85 and 1.39 for error rate of 10% and 1%, respectively. The performance trends in this scenario are observed to be radically different from those in the random error case (Fig. 9.11). Throughput performance of TCP Vegas reduces by about 80% (compared to the uniform error case) when the error rate is 2%.

SACK TCP now performs better than TCP New-Reno over a range of error rates from 0% to 5%.

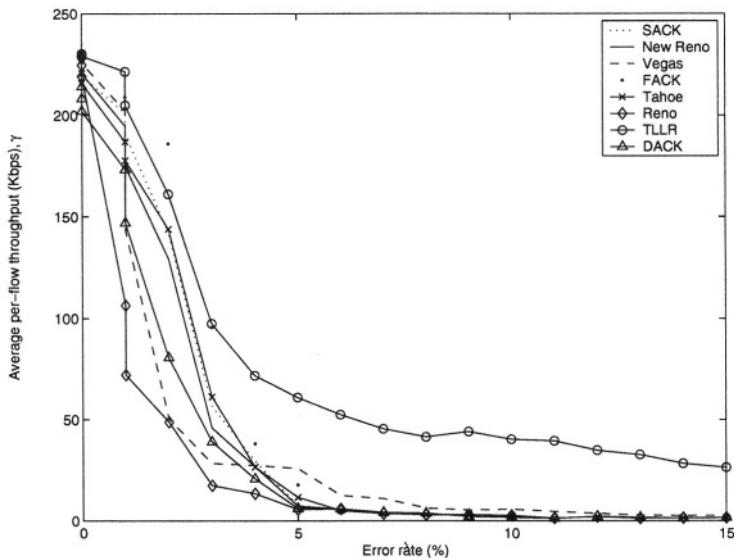


Figure 9.11. Throughput performances for the different TCP variants (scenario 5).

Since the system dynamics is now characterized by a large number of timeouts, high ratio of the number of multiple (three or more) packet drops to the number of single packet drops in a single window and loss of a large number of acknowledgements, performance of TCP Tahoe becomes comparable to the performance of SACK TCP. Also, we observe that the performance of TCP New-Reno with TCP-aware link level retransmission deteriorates considerably compared to that in the uniform error case.

TCP Reno is observed to perform the worst in this correlated error scenario. Although the number of fast retransmits is fewer than that in the uniform error case, most of the fast retransmits are due to multiple packet drops (three or more) in a single window. For example, with 1% error rate, there are 402 fast retransmits, among which 375 are due to multiple drops (Table 9.4). Again, 327 of the 375 packet drops are due to three or more packet drops in a single window. Most of these multiple (more than three) drop cases are observed to be followed by timeouts. In fact, when three or more packets are dropped from a window of data, the Reno sender is forced to wait for a timeout most of the time.

TCP Tahoe is observed to perform significantly better than Reno under correlated error scenarios. Since after switching to *slow-start*, Tahoe

Table 9.4. Fast retransmit/fast recovery analysis under correlated error rate 1% and 10% in 1000 second run

| TCP variant | TO | FR | MD | SND | FLT |
|------------------|-------------|-----|-----|------|-------|
| Error rate = 1% | | | | | |
| Tahoe | 8 | 539 | 538 | 0 | 14.98 |
| Reno | 431 | 402 | 375 | 226 | 13.01 |
| New-Reno | 7 | 544 | 541 | 1 | 14.99 |
| SACK | 4 | 553 | 547 | 1205 | 14.97 |
| DACK | 12 | 530 | 484 | 260 | 14.42 |
| FACK | 0 | 548 | 546 | 3284 | 14.93 |
| TLLR | 0 | 0 | 0 | | |
| Vegas | 157 | 445 | | | 14.20 |
| Error rate = 3% | | | | | |
| Tahoe | 195 | 214 | 210 | 59 | 14.42 |
| Reno | 346 | 95 | 93 | 25 | 12.80 |
| New-Reno | 230 | 234 | 231 | 164 | 13.93 |
| SACK | 330 | 251 | 246 | 870 | 14.02 |
| DACK | 254 | 215 | 210 | 103 | 11.82 |
| FACK | 159 | 349 | 348 | 1941 | 14.51 |
| TLLR | 0 | 0 | 0 | 0 | 0 |
| Vegas | 381 175E | 164 | | | 10.82 |
| Error rate = 10% | | | | | |
| Tahoe | 140 | 0 | 0 | | |
| Reno | 140 | 0 | 0 | | |
| New-Reno | 140 | 0 | 0 | | |
| SACK | 140 | 0 | 0 | | |
| DACK | 145 | 0 | 0 | | |
| FACK | 140 | 0 | 0 | | |
| TLLR | 68 | 0 | 0 | | |
| Vegas | 248 0E | 0 | | | |

forgets all outstanding data that were transmitted earlier and increases its congestion window upon receipt of each acknowledgement, it results in multiple transmission attempts for same packets (some of which have presumably been lost already), and consequently, the throughput improves for the multiple drop cases.

Since the number of packet drops per transmission window increases in the correlated error case, SACK TCP performs better than New-Reno. TCP New-Reno requires $n \times RTT$ to recover from n packet drops in a single window whereas SACK recovers much faster. This is mainly due to its better pipe estimation method and usage of a larger retransmission pool. For the same reason FACK TCP performs even better than SACK TCP.

Since a large number of acknowledgements are lost due to correlated channel error, sluggishness in transmitting acknowledgements in the case of TCP New-Reno with DACK option does not significantly impact the throughput performance. For the same reason, sharp response to packet losses in the case of TCP Vegas is not very conducive to improving TCP throughput performance in the correlated error case. Under different channel error rates and channel error correlation, the fixed values of α and β do not work well. Also, TCP Vegas experiences the highest number of timeouts.

For error rate greater than 7%, the energy consumption is observed to be more or less same for all the TCP variants (except for Vegas and New-Reno with TLLR) (Fig. 9.12). This is due to the fact that for a large error rate all of the TCP variants experience almost the same number of timeouts and there may be no *fast retransmit* at all.

3.3 Summary of the Results

Throughput, fairness and energy performances of the different basic TCP variants have been investigated in a wide-area cellular wireless environment for both uniform and correlated wireless channel errors. The following provides a summary of the key observations:

- Implications of the fast retransmit and timeout events on the TCP performance largely depend on the wireless channel error characteristics (e.g., error rate and degree of channel error correlation). For example, in the case of TCP Vegas, sharp response of the sender due to its fine-grained expiry mechanism works well under uniform error case while the same strategy results in poor performance under correlated error scenarios. Therefore, although coarse timeout is undesirable under random losses, the ineffectiveness of the ‘Vegas expiry’ mechanism under correlated error cases

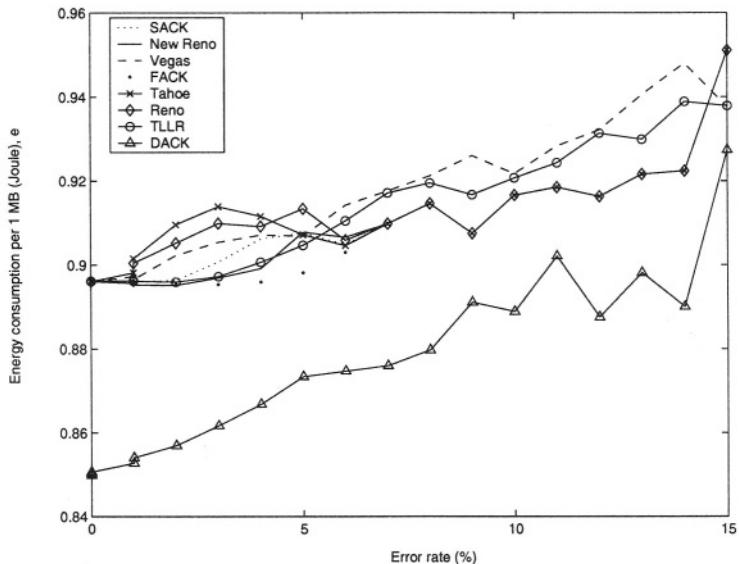


Figure 9.12. Energy consumption for the different TCP variants (scenario 5).

suggests that very sharp timeouts may also not be desirable under such error conditions.

- A method which provides better estimation for the number of outstanding segments is always conducive to the TCP throughput performance under both random and correlated error scenarios. Due to this reason, FACK TCP is observed to be consistently better than SACK TCP and TCP New-Reno under both random and correlated error scenarios.
- Since the impact of timeout becomes more dominant, the throughput performances of all of the basic end-to-end TCP variants (which primarily differ in the *fast recovery* mechanism) suffer under correlated error scenarios. Again, end-to-end protocols with link level retransmissions (e.g., snoop protocol) also suffer serious performance degradation in case of correlated channel errors.
- Lower degree of wireless channel error correlation is more conducive to energy saving at the mobile TCP sinks.
- In a wide-area wireless scenario, TCP connections with high bandwidth-delay product may not get their fair share. Splitting a single TCP flow (with a high bandwidth-delay product) to multiple flows (say,

x flows) increases the throughput approximately x times for a certain range of low error rates.

3.4 Outlook

- TCP performance in a wired-cum-wireless environment can be improved significantly by using some transport level mechanism to differentiate between the congestion loss and wireless channel loss and then adjusting the window adaptation mechanism accordingly.
- The transmission window size at the TCP sender should be adapted differently depending on the degree of correlation in the wireless link errors. In addition, timer granularity may be adapted dynamically based on error correlation.
- The throughput performance of an end-to-end TCP designed for a wide-area cellular environment based on transport level modifications should be compared against the maximum achievable throughput envelope. This envelope, which can be characterized empirically based on simulation results, defines the maximum throughput achievable under any TCP-like window-based end-to-end transmission control mechanism.

4. Maximum Achievable TCP Throughput Under Window-Based End-to-End Transmission Control

In this section, we obtain the upper bound on the TCP throughput performance under window-based end-to-end transmission control (i.e., throughput of *ideal end-to-end TCP (IE²-TCP)*) and compare it against the throughput performances of TCP New-Reno, SACK TCP and FACK TCP. For any window-based transmission control, the timeout mechanism is a must. Note that, all of the basic TCP variants use similar timeout mechanism along with exponential RTO backoff. Also, they use the *fast retransmit* mechanism (i.e., immediate retransmission of lost segment detected by *tcp_rexmit_thresh* number of duplicate ACKs). However, the basic end-to-end TCP variants mainly differ in the implementation of the *fast recovery* mechanism. Note that, for the maximum possible throughput performance in a wired-cum-wireless scenario, *IE²-TCP* should be aggressive enough in retransmission and in measuring the outstanding data segments. The upper bound on the throughput performance (achievable by the *IE²-TCP*) can be obtained using the simulation model described before based on the following assumptions:

- IE^2 -TCP uses the timeout mechanism along with exponential RTO backoff and it uses *fast retransmit* mechanism to retransmit a lost segment after reception of *tcp-rexmit-thresh* number of duplicate ACKs.
- IE^2 -TCP retransmits a lost segment after *tcp-rexmit-thresh* number of partial duplicate ACKs (in order to avoid some possible timeouts by retransmissions during *fast recovery*).
- Rather than using a go-back-N type of retransmission, IE^2 -TCP uses selective acknowledgement (SACK)-based retransmission.
- IE^2 -TCP uses the information in the SACK acknowledgement to estimate outstanding data.
- IE^2 -TCP maintains the transmission window size equal to the receiver-advertized window size all the time.

Figs. 9.13-9.14 show some typical results on the long-term average throughput performance of IE^2 -TCP when compared to those of SACK TCP, TCP New-Reno and FACK TCP under both random error and correlated error conditions. As is evident from Fig. 9.13, under random packet error in the wireless link IE^2 -TCP can achieve a throughput which is higher by as much as 200% (e.g., for packet error rate of 4%) than that due to SACK TCP and FACK TCP. Improvement in throughput performance can be even higher under correlated error scenarios (e.g., 250% improvement for packet error rate of 4% with average error burst length of 3.85) (Fig. 9.14). Therefore, it can be concluded that the basic TCP variants perform significantly poor compared to the *ideal TCP* (i.e., IE^2 -TCP) in a wired-cum-wireless scenario. Since the throughput performance of IE^2 -TCP characterizes the envelope of the maximum possible throughput under a TCP-like window-based end-to-end transmission control, the TCP modifications based on end-to-end approaches should use this as the benchmark performance.

5. Modifications to the Basic TCP Variants for Wireless Networks

5.1 Sender-Only Modifications

- *K-SACK* [15]: *K-SACK* is an efficient SACK-TCP variant of TCP New-Reno that requires modification at the TCP sender only. The novelty of this protocol is that it differentiates between wireless losses and congestion losses from anticipated loss pattern and behaves accordingly. During *fast recovery*, in case of wireless loss, it

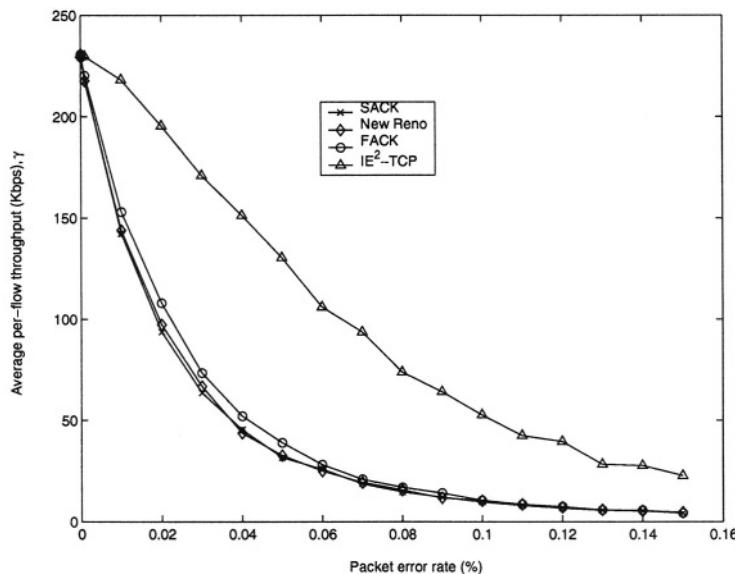


Figure 9.13. Throughput performances for SACK TCP, TCP New-Reno, FACK TCP and IE^2 -TCP under random packet error in the wireless link.

does not halve the congestion window instead it stalls the growth of the window. The *fast recovery* phase is partitioned into two phases: *halt growth* phase and *K-recovery* phase. During *K-recovery* phase, the sender apprehends congestion only if it anticipates *K lookahead-loss* within a loss window of size *lwnd*, where *K* and *lwnd* are appropriately chosen parameters for the protocol.

The parameter *lwnd* can be chosen to be equal to the number of outstanding packets. The parameter *lookahead-loss* estimates the number of losses in the outstanding packets. For each such loss the sender has received either (i) at least *max_dupack* number of duplicate acknowledgements, or (ii) *max_dupsack* number of selective acknowledgements with higher sequence numbers. For lookahead-loss below *K* (e.g., *K* = 2), the sender assumes loss due to channel error.

Upon entering into the *fast recovery* phase, if lookahead-loss is less than *K*, the sender enters into the *halt growth* phase, otherwise to the *K-recovery* phase and transition between these phases occurs depending on the value of the lookahead-loss.

During the *halt growth* phase, congestion window remains frozen and the sender remains in this phase until the lookahead-loss be-

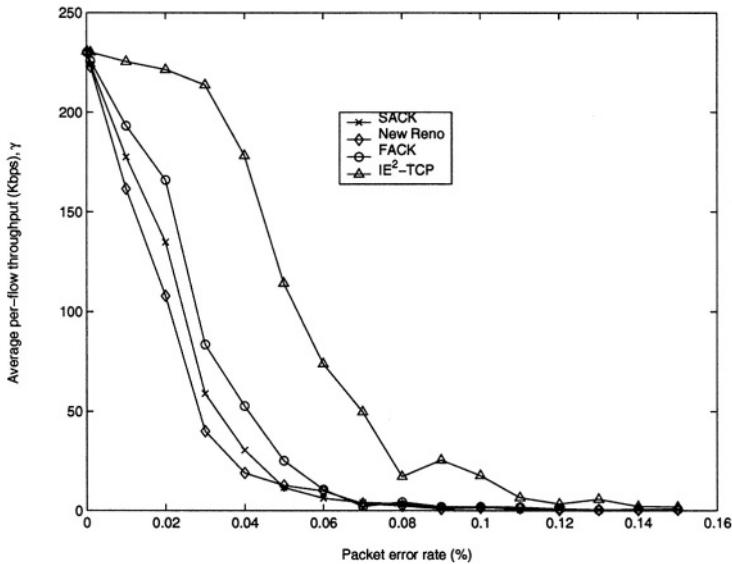


Figure 9.14. Throughput performances for SACK TCP, TCP New-Reno, FACK TCP and IE^2 -TCP under correlated packet error in the wireless link (for mobile velocity of 3 km/hr).

comes at least K . When the sender enters into the K -recovery phase for the first time, $ssthresh$ is set to $\max(cwnd/2, 2 \times MSS)$ and $cwnd$ is set to $ssthresh$. Upon re-entry, $ssthresh$ remains unchanged, but congestion window is set to $\max(cwnd/2, 1 \times MSS)$. This may cause the congestion window to change multiple times during the same *fast recovery* period. As a result, upon exiting *fast recovery*, depending on the value of $ssthresh$ and $cwnd$, K -SACK TCP may find itself in the *slow start* phase instead of the *congestion avoidance* phase.

K -SACK performs a better *pipe* (i.e., number of packets in the link) estimation by measuring the *pipe* as the number of unacknowledged segments which are not marked *lost* or marked *lost* and retransmitted. This improved pipe estimation results in faster error recovery.

The main advantage of K -SACK is that it detects packet loss due to wireless channel error without explicit notification and it can be used seamlessly in wired or wired-cum-wireless environment without separate tuning. It was observed to perform very well under high *i.i.d.* error condition. Simulation results showed that,

at 5% *i.i.d.* error rate the throughput was almost twice of that due to SACK-TCP over New Reno.

Although *K*-SACK performs well under random packet errors in the wireless channel, the performance of this scheme will suffer under correlated error conditions. This is due to the fact that, a moderate degree of correlated error will increase lookahead-loss to *K* and thus push *K*-SACK into *K*-recovery phase. Therefore, during *fast recovery* the sender may spend significant amount of time in *K-recovery* phase compared to that in the *halt growth* phase.

To combat the situation due to correlated packet loss in the wireless channel, higher value of *K* can be chosen. However, it will increase the chance of congestion collapse in the wired Internet. Since in a real Internet scenario, only a small percentage of packet losses (e.g., 15% [16]) is recovered by *fast recovery*, enhancement only in the *fast recovery* mechanism as in *K*-SACK is not likely to improve the TCP performance significantly compared to the other TCP variants in a wide-area wireless Internet. In addition, under pathological scenarios *K*-SACK would suffer from oscillation between the *K-recovery* and the *halt growth* phase which would reduce *cwnd* to small values, and consequently, the performance may fall below that of SACK TCP.

- *TCP Westwood (TCPW)* [17]: TCPW uses a sender-side modification of TCP congestion window adaptation algorithm which is based on estimation of the bandwidth used by the TCP connection via monitoring the rate of returning ACKs. The bandwidth estimate (BWE) is used to set the congestion window and the slow-start threshold after a congestion episode (i.e., after three duplicate acknowledgements or after a timeout), and thereby, TCPW achieves a “faster” error recovery. The basic congestion window dynamics during *slow-start* and *congestion avoidance* are unchanged. For bandwidth estimation, TCPW uses a discrete-time exponential filter with variable coefficients as shown in (9.2) below

$$\hat{b}_k = \alpha_k \hat{b}_{k-1} + (1 - \alpha_k) \left(\frac{b_k + b_{k-1}}{2} \right) \quad (9.2)$$

where \hat{b}_k is the filtered estimate of the available bandwidth at time $t = t_k$, $\alpha_k = (2\tau - \Delta_k)/(2\tau + \Delta_k)$, $\Delta_k = t_k - t_{k-1}$ and $1/\tau$ is the cutoff frequency of the filter. When the interarrival time Δ_k increases, the last value of \hat{b}_{k-1} has less significance, while a recent \hat{b}_{k-1} is given higher significance. The coefficient α_k decreases when the interarrival time increases, and thus the previous value of b_{k-1}

has less significance with respect to the last two samples which are weighted by $(1 - \alpha_k)/2$. Using the estimated bandwidth, the values of *ssthresh* and *cwnd* are updated as follows:

```
/* algo. for n dup acks */

if (n dupacks are received)
    ssthresh = (BWE * RTTmin) / seg_size;
    if (cwnd > ssthresh)
        cwnd = ssthresh
    endif
endif

/* algo. for coarse timeout expiration */

if (coarse timeout expires)
    ssthresh = (BWE * RTTmin) / seg_size;
    if (ssthresh < 2)
        ssthresh = 2;
    endif;
    cwnd = 1;
endif
```

As an end-to-end solution to error and congestion control in mixed wired and wireless networks, TCPW was observed to provide significant throughput gain over TCP Reno and SACK TCP for both random and correlated error conditions in the wireless link. Also, the observed throughput fairness among TCPW flows across different round trip times was better compared to that for TCP Reno flows. This was due to the fact that, under TCPW long connections suffered less reduction in *cwnd* and *ssthresh*.

As the round trip time increases, the time for the feedback information to reach the TCP sender also increases which reduces the effectiveness of TCPW. Also, TCPW performs poorly when random packet loss rate exceeds a small threshold. During long correlated error burst, TCP timeouts occur which affect TCPW performance severely.

- *TCP Santa Cruz* [18]: TCP Santa Cruz (SC) uses new congestion control and error recovery strategies which are designed to work with path asymmetries, networks with lossy links, limited

bandwidth, variable delay and out-of-order packet delivery. Since congestion losses are preceded by an increase in the network bottleneck queue (which may not be true for random losses in the wireless link), TCP-SC monitors the queue developing over a bottleneck link and determines whether congestion is incipient in the network and responds by increasing or decreasing the congestion window accordingly. The goals are to perform timely and efficient early retransmission of any lost packet, eliminate unnecessary retransmissions for correctly received packets when multiple losses occur within a window of data, and provide RTT estimates during periods of congestion and retransmission.

The congestion control algorithm in TCP-SC is based upon the measurement of *relative delay* that packets experience with respect to each other as packets are transmitted through the network. For each transmitted packet, TCP-SC sender maintains information about the transmission time of the packet, the arrival time of an ACK for that packet and arrival time of the data packet at the receiver (which is reported by the receiver in its ACK). Then for any two data packets i and j , the *relative forward delay* $D_{j,i}^F$ can be calculated as $D_{j,i}^F = R_{j,i} - S_{j,i}$, where $D_{j,i}^F$ represents the additional forward delay experienced by packet j with respect to packet i , $S_{j,i}$ is the time interval between the transmission of the packets, and $R_{j,i}$ is the interarrival time of data packets at the receiver. Based upon the values of $D_{j,i}^F$ over time, the change in the states of the queues (and specially the bottleneck queue) are determined. For example, if the sum of relative delays over an interval is 0, it implies that no additional congestion or queueing is present in the network at the end of the interval with respect to the beginning.

If n is the operating point of the congestion control algorithm (i.e., n is the desired number of packets in the bottleneck queue which is assumed to be one packet more than the BDP of the network), it attempts to maintain the total number of packets queued at the bottleneck link from the beginning of the connection until t_i (N_{t_i}) to n , where $N_{t_i} = N_{t_{i-1}} + M_{W_{i-1}}$, with $M_{W_{i-1}}$ being the additional amount of queueing introduced over the previous window W_{i-1} , and $N_{t_1} = M_{W_0}$. Now $M_{W_{i-1}}$ can be calculated based on the relative delay measurements as follows: $M_{W_{i-1}} = \frac{\sum D_k^F}{\bar{t}_{pkt}}$ where \bar{t}_{pkt} is the average packet service time and k is the number of packet pairs within window W_{i-1} . Note that, the average packet service time can be calculated based on the timestamps returned by the receiver.

TCP-SC adjusts the congestion window once for each interval equal to the amount of time it takes to transmit one window of data. Over this interval, $M_{W_{t_{i-1}}}$ is calculated and at the end of the interval it is added to N_{t_i} . If $N_{t_i} < n - \delta$ (δ is some fraction of a packet), the congestion window is increased linearly, while if $N_{t_i} > n + \delta$, the congestion window is decreased linearly during the next interval, otherwise the congestion window is maintained at its current size. In this way, since adjustment of the congestion window does not depend on the arrival of ACKs, the congestion control algorithm becomes robust to ACK loss.

TCP-SC sender receives precise information on each packet correctly received and it uses a tighter estimate of the RTT per packet. TCP-SC recovers losses quickly without necessarily waiting for three duplicate acknowledgements from the receiver. Let packet i initially transmitted at time t_i is lost and marked as a hole in the ACK window. Packet i is retransmitted as soon as an ACK arrives for any packet transmitted at time t_x such that $t_x > t_i$, and $t_{current} - t_i > RTT_e$, where $t_{current}$ is the current time and RTT_e is the estimated round trip time of the connection.

TCP-SC was observed to provide high throughput and low end-to-end delay and delay variance over networks with a simple bottleneck link, networks with congestion in the reverse path of the connection, and networks which exhibit path asymmetry. TCP-SC can be implemented as a TCP option by utilizing the extra 40 bytes available in the options field of the TCP header.

TCP-SC does not differentiate between packet loss due to congestion and packet loss due to wireless link error. In fact, effectiveness of the proposed relative delay-based approach under correlated wireless channel losses is unclear.

- *Selective Slow Start (SSS)* [19]: Under this scheme, TCP attempts to distinguish between packet loss due to congestion and packet loss due to handoff based on the pattern of losses. For this, after a timeout the TCP sender counts the number of lost segments in the last S attempts including the current one and compares it with a pre-calculated $LIMIT$. If the number of lost segments is higher than $LIMIT$, the sender assumes that these losses are due to network congestion and initiates *slow start*, otherwise it continues to transmit at its current rate using the same timer values assuming that the losses are due to handoff. The value of the parameter $LIMIT$ is set in a way that it can account for all segment losses during a handoff. For example, $LIMIT$ can be set as $LIMIT =$

$\frac{K \times C \times T_h}{Seg}$, where C is the senders transmission rate (in bps), T_h is the average handoff duration (in $seconds$), Seg is the average segment size and K is a safety factor.

Although the SSS scheme was shown to perform better than standard TCP in wireless ATM environment for high handoff rate (e.g., 0.2/*second*), the main limitation of this scheme is that for rate adaptation at the TCP sender it does not take wireless link conditions into account. Also, using a high value of *LIMIT* may cause congestion collapse in the wired Internet.

5.2 Receiver-Only Modifications

- *Freeze-TCP* [20]: This is a proactive mechanism in which the receiver notifies the sender of any impending ‘blackout’ situation (e.g., due to wireless channel fading or handoff) by ‘zero window advertisement (ZWA)’ and prevents the sender from entering into congestion avoidance phase. Upon receiving ZWA, the sender enters into the ‘zero window probes (ZWP)’ mode and freezes corresponding timers. While in ZWP mode, the sender transmits zero window probes and the interval between successive probes grows exponentially until it reaches 1 minute, where it remains constant. When the ‘blackout’ period is over, the receiver sends TR-ACKs (Triplicate Reconnection ACKs) for the last data segment successfully received to enable *fast retransmit* at the TCP sender.

Ideally, the ‘warning period’ prior to disconnection (i.e., how much in advance should the receiver start ZWA) should be long enough to ensure that exactly one ZWA gets across the sender. If the warning period is longer than this, the sender will be forced into ZWP mode prematurely resulting in idle time prior to disconnection. If the warning period is too small, the receiver might not have enough time to send out a ZWA which will cause the sender’s congestion window to drop.

One drawback of this approach is that, the receiver needs to predict the impending disconnections. For this, some cross-layer information exchanges may be necessary.

- *ACK-Fragmentation Protocol* [21]: If the mobile host, after sending three duplicate acknowledgements, receives a new retransmitted packet from the sender, it sends N equally spaced cumulative acknowledgements for this new packet within one round-trip time. If for this received packet the sequence number of the first octet is S and last octet is F , the i th cumulative acknowledgement will

acknowledge upto octet $S + i(F - S)/N$. The value of N can be computed from the random packet loss probability in the wireless link, the link capacity, round-trip time and buffer size at the bottleneck link [21].

The effectiveness of this protocol comes from the fact that increased number of acknowledgements within same round-trip time expedites the resumption of normal transmission window at the TCP sender. Moreover, it increases the chance that the sender receives at least one acknowledgement.

5.3 Modifications at the Sender and the Receiver

- *Wireless TCP (WTCP)* [22]: WTCP is an end-to-end reliable transport layer protocol for wired-cum-wireless networks which uses *rate-based* transmission control instead of window-based transmission control and the rate adaptation computations are performed by the receiver. Since the receiver performs the rate computations, the effects of delays variations and losses in the ACK path are eliminated. The goal of WTCP is to decrease the transmission rate aggressively in case of congestion (so that the congestion alleviates quickly) and decrease less aggressively in the case of incipient congestion to improve efficiency.

To distinguish between congestion-related and non-congestion-related packet loss, the WTCP receiver, using a history of packet losses, computes the mean and mean deviation in the number of packet losses when the network is predicted to be uncongested. If the loss is predicted to be due to congestion, the transmission rate is decreased aggressively by 50%. For non-congestion related packet loss WTCP uses the ratio of the average interpacket delay observed at the receiver to the interpacket delay at the sender (rather than using packet loss and retransmission timeouts) to control transmission rate. The WTCP receiver maintains two state variables *lavg_ratio* and *savg_ratio* for the long-term and short-term running averages of the ratio of the observed sending rate at the receiver to the actual sending rate at the sender. At any time, the receiver can be in one of the three states—*increase* (when $lavg_ratio > \alpha_+$ and $savg_ratio > \beta_+$), *decrease* (when $lavg_ratio > \alpha_-$ and $savg_ratio > \beta_-$), *maintain*.

WTCP uses SACK and no retransmission timer for loss recovery. The sender tunes the desired rate for ACK transmission by the

receiver so that it receives at least one ACK in a threshold period of time and can react to the new transmission rate.

In summary, WTCP is significantly different from TCP and tuning of the different parameters would be required to optimize the performance of WTCP.

- *Explicit Congestion Notification (ECN) TCP* [23]: This proposal is based on the TCP with explicit congestion notification [24] mechanism where the routers should inform the TCP sender of incipient congestion so that the TCP senders can lower the transmission rate. For example, if RED (Random Early Detection) mechanism is used, a router signals incipient congestion to TCP by setting the Congestion Experienced (CE) bit in the IP header of the data packet when the average queue size lies between two thresholds min_{th} and max_{th} . In response, the receiver sets the ‘ECN-Echo (ECE)’ bit in the ACKs and when the sender receives an ACK packet with ECE bit set, it invokes the congestion control mechanism. In case of packet loss in the wireless channel, the receiver sends ACK packets without CE bit set and on receiving three duplicate ACKs without CE bit set the sender deduces that the packet loss is due to errors in the wireless channel. In this case, the sender does not invoke the *congestion avoidance* algorithm.

In case of long outage in the wireless channel, this scheme may not work well. Again, this can be used only for ECN-capable TCP.

- *Eifel Algorithm* [25]: A large number of timeouts and fast retransmissions that occur at a TCP sender have been observed to be *spurious*. That is, the ACKs received at the sender after a timeout or fast retransmit are actually the ACKs for the original packets and not for the retransmitted packets. Spurious fast retransmits result when the packet reordering in the Internet results in a reordering length more than the duplicate ACK threshold. Due to spurious timeouts and spurious retransmissions, the TCP sender unnecessarily reduces the transmission rate and enters into a go-back-N type of retransmission mode. When the sender enters into the retransmission mode, it causes the receiver to send duplicate ACKs which may result in *fast retransmit* after a timeout resulting in further reduction in transmission rate. The *Eifel algorithm* was proposed to eliminate the problems caused by *spurious timeouts* and *spurious fast retransmits* in TCP.

The Eifel algorithm uses the TCP timestamp option. Each packet transmitted by the sender is timestamped and the values of *ssthresh*

and $cwnd$ are also stored. The receiver returns the timestamp value of the packet in the corresponding ACK. After a timeout or a fast retransmit, the values of $ssthresh$ and $cwnd$ are updated according to the usual algorithms and the packet is retransmitted. After receiving the ACK, if the sender finds that it is for the original packet, it restores the values of $ssthresh$ and $cwnd$.

This algorithm does not take packet loss due to wireless channel error into account, and therefore, does not solve the problems of TCP in a wired-cum-wireless network.

5.4 Modifications at Base Station: Non-End-to-End Transport Level Approaches

- *Indirect-TCP (I-TCP)* [27]: I-TCP splits a transport layer connection between the MH and the FH into two separate connections – one over the wireless link between the MH and its mobile support router (MSR) at the base station and the other between the MSR and the FH over the fixed network. After receiving the data packets destined to the MH, the MSR sends ACKs to the FH and forwards the packets to the MH using a separate transport protocol designed for better performance over wireless links. Such a transport protocol for the wireless link can support notification of events such as disconnections to the link-aware and mobility-aware applications. In this case, the mobile hosts can run simple wireless access protocol to communicate with the MSR and the MSR manages the communication overheads for communicating with the FHs. For example, when an MH wants to communicate with a FH, it sends a request to its current MSR, and the MSR then establishes a TCP connection with the FH. Whenever an MH moves to another cell, the entire connection needs to be moved to the new MSR which should be completely transparent to the fixed network.

Performance of I-TCP was observed to be better than regular TCP in wide-area wireless networks. However, end-to-end TCP semantics are not maintained in this protocol and applications running on the mobile host have to be relinked with the I-TCP library. Also, I-TCP is not suitable for cases where the wireless link is not the last part of a connection path, because in such a case a particular connection may need to be split several times resulting in performance degradation.

- *Using ICMP Control Messages* [29]: A scheme for improving TCP performance in wireless networks using ICMP control messages was proposed in [29]. In this scheme, the BS transmits an ICMP-DEFER message to the FH when it fails in its first attempt to transmit the packet to the MH over the wireless link. On receiving the ICMP-DEFER message for a particular packet, the sender resets the corresponding timer. When all local retransmissions fail, the BS notifies the sender using ICMP-RETRANSMIT message so that the sender can retransmit the packet.
- *Using Multiple Acknowledgements* [30]: The scheme proposed in [30] maintains end-to-end semantics and uses two types of ACKs, namely, *partial ACK* (ACK_p) and *complete ACK* (ACK_c) to distinguish losses due to congestion from losses due to wireless link error. The sender uses the regular TCP mechanism with *slow-start*, *congestion avoidance*, *fast retransmit* and *fast recovery*. The complete acknowledgement ACK_c is sent by the receiver and the partial acknowledgement ACK_p is sent by the BS. ACK_p with sequence number N_a informs the sender that packet(s) with sequence numbers up to $N_a - 1$ have been received by the BS, and the BS is facing problems in forwarding the packets with sequence numbers from ACK_c up to $N_a - 1$ (i.e., it has not received the ACK from the wireless host even after waiting for the maximum possible delay time between the transmission of a packet from BS and reception of the ACK from the mobile host). Then the sender marks the corresponding packets and updates RTO to give the BS more time to perform retransmission. If timeout occurs, and the packets are still marked, the sender will backoff the timer only without invoking any congestion control method. On receipt of ACK_c , the sender works similar to a normal TCP sender.

When the BS receives an in-sequence packet, or an out of sequence packet which is not in its buffer, the packet is buffered and transmitted to the mobile host. The BS starts a timer with the timeout value equal to the maximum time which can elapse between the reception of a packet at the BS and its acknowledgement from the mobile host. If the BS receives an out of sequence packet which is present in the buffer, it sends an ACK_p to the sender. If the local timer at the BS expires before receiving an acknowledgement from the mobile host, the BS retransmits the packets and sends an ACK_p to the sender.

- *Mobile-End Transport Protocol (METP)* [31]: METP hides the wireless link from the rest of the Internet in a wired-cum-wireless

network by terminating the TCP connection at the BS on behalf of mobile host. TCP/IP between the BS and a mobile host is replaced by a low overhead protocol. All TCP connections are handled by METP at the BS which negotiates with another host in the Internet to open or close a TCP connection, and keeps the connection state and sending and receiving buffers. For data transfer from a mobile host to an Internet host through a TCP connection, the mobile host sends the data to the BS, and METP sends them out as TCP segments to the destination. When a TCP/IP packet destined to the mobile host arrives at the BS, METP sends an ACK and puts it in the receiving buffer from where a separate process transmits it to the mobile host. In case of temporary link failure, since METP at the BS continues to receive data from the fixed host, it sends out ACK with smaller advertised window. The advertised window is retained to its original level when the wireless link becomes good again.

METP uses link layer retransmissions and acknowledgements for reliable data delivery over the wireless link. If no acknowledgement is received immediately after a data frame transmission, the frame is retransmitted after a random backoff interval. To avoid buffer overflow, flow control is achieved by the METP at the receiver through periodically sending out a feedback packet to inform the sender of the available buffer space.

METP uses a header size of only 12 bytes for packets exchanged between the mobile host and the BS. The header size can be further reduced by using the idea of header compression. In case of handoff the old BS opens a separate TCP connection with the new BS and sends all data and state information.

The merits of this protocol are that by using small header it incurs less overhead for wireless transmission and it avoids activation of the TCP congestion control mechanisms in case of packet loss in the wireless channel. METP was observed to provide better throughput performance than split-connection TCP and split-connection TCP with selective ACK under varying handoff interval and BS buffer size. However, METP does not maintain the strict end-to-end semantics and suffers from the scalability and reliability problems (due to failure of the BS).

- **WTCP [32]:** WTCP uses a modified flow and error control protocol between the BS and the mobile host. After receiving a TCP/IP packet from a fixed host, WTCP at the BS buffers it (if this is the next packet expected from the fixed host or if the packet has a

larger sequence number than expected) along with its arrival time. WTCP then transmits the buffered packet to the mobile host and when a packet is transmitted to the mobile host, the BS schedules a new timeout if there is no other timeout pending. Upon receiving the ACK from the mobile host, WTCP frees the corresponding buffer and sends the ACK to the fixed host. WTCP performs local error recovery based on duplicate acknowledgements or timeout. In case of timeout, WTCP reduces the transmission window size to one assuming subsequent bad channel condition. As soon as the BS receives an ACK, the transmission window size is set again to the receiver advertised window size. Also, upon reception of a duplicate ACK, WTCP opens the transmission window in full.

Another feature of WTCP is, it attempts to hide the time spent by the BS for local error recovery by adding that time to the timestamp value of the corresponding segment so that the TCP's round trip time estimation at the source is not affected. In this way, the TCP source's ability to effectively detect congestion in the wired network is not impacted.

Due to its aggressiveness in window adaptation, WTCP was observed to provide better/equal throughput performance compared to I-TCP and *snoop TCP*. However, a more optimized flow and congestion control scheme along with header compression (as in METP) can be used under similar scenario to provide even better performance.

A qualitative comparison among the different end-to-end TCP modifications is provided in Table 9.5.

6. Chapter Summary

Performances of the different basic TCP variants (e.g., TCP Tahoe, TCP Reno, TCP New-Reno, SACK TCP, FACK TCP, TCP Vegas) have been analyzed in a wide-area cellular network by using *ns-2-based* computer simulations. Impact of the wireless channel error characteristics on the slow-start, congestion avoidance and timeout mechanisms of TCP has been explored by analyzing the transport level system dynamics. The maximum achievable end-to-end throughput for a TCP-like window-based end-to-end transmission control mechanism has been evaluated which can serve as a benchmark for performance evaluation of any transport protocol designed (to enhance TCP performance in wide-area wireless networks) based on transport level modifications to the basic TCP mechanisms. A brief overview of the different proposed

Table 9.5. Comparison among the transport-level TCP modifications.

| TCP Variant | Maintains end-to-end semantics | Can handle end-to-end encryption | Requires intermediaries' support | Can handle random packet loss | Can handle burst errors efficiently |
|---------------------|--------------------------------|----------------------------------|----------------------------------|-------------------------------|-------------------------------------|
| K-SACK | Yes | Yes | No | Yes | No |
| Westwood | Yes | Yes | No | Yes | No |
| Santa Cruz | Yes | Yes | No | Yes | - |
| SSS | Yes | Yes | No | Yes | No |
| Freeze-TCP | Yes | Yes | No | Yes | Yes |
| ACK-fragmentation | Yes | Yes | No | Yes | No |
| WTCP (rate-based) | Yes | Yes | No | Yes | Yes |
| ECN | Yes | No | Yes | Yes | No |
| Eifel | Yes | Yes | No | No | No |
| I-TCP | No | No | Yes | Yes | No |
| Using ICMP | No | No | Yes | Yes | Yes |
| Multiple-ACK | Yes | No | Yes | Yes | Yes |
| METP | No | No | Yes | Yes | Yes |
| WTCP (window-based) | Yes | No | Yes | Yes | Yes |

TCP mechanisms (based on transport level modifications) along with a qualitative performance comparison have also been presented.

Future research in end-to-end TCP design for wired-cum-wireless networks should address

- how to distinguish among packet losses due to wireless channel error and congestion error (e.g., by using some estimation/filtering mechanisms) and comparison among the different (estimation/filtering) mechanisms
- how to exploit the above information for adjusting the transmission rate at the sender in the case of packet loss to design end-to-end TCP based on modifications at the sender only
- fairness, energy-efficiency and TCP-friendliness of an end-to-end TCP customized for wired-cum-wireless networks.

Acknowledgments

The authors acknowledge the support from Telecommunications Research Labs (*TRLabs*), Winnipeg, Canada, for this research.

Notes

1. In the Internet community, both NPDU and TPDU are known as *packet* even though in OSI terminology a *segment* is referred to as a TPDU. The terms ‘packet’ and ‘segment’ are used interchangeably in this article.
2. The terms ‘flow’ and ‘connection’ are used interchangeably in this article.
3. The bandwidth-delay product for a TCP connection refers to the product of the round trip delay (T) for the connection and the capacity of the bottleneck link (μ) in its path.
4. This refers to a new acknowledgement received during *fast recovery* which acknowledges some but not all of the packets that were outstanding at the start of the *fast recovery* phase.
5. MSS refers to the *Maximum Segment Size*.
6. Timeout caused by this fine-grained timer is referred to as ‘Vegas expiry’.
7. Here, $f_d = v/\lambda$ = mobile speed/carrier wavelength. The value of t_{tcp} determines the minimum fade duration.
8. It refers to the maximum fading attenuation which still allows correct reception of a packet.
9. This is similar to the fairness function used in [50] to quantify the fairness in a shared resource system with n users: $F = \frac{\left(\sum_{i=1}^n x_i\right)^2}{n \sum_{i=1}^n x_i^2}$ (where x_i is the i th user’s throughput).

References

- [1] Iren, S., and P. D. Amer, “The transport layer: Tutorial and survey,” *ACM Computer Surveys*, vol. 31, no. 4, pp. 360-405, Dec. 1999.

- [2] Ayala, R., K. Basu, and S. Elliot, "Internet technology based infrastructure for mobile multimedia services," *Proc. IEEE WCNC'99*, pp. 109-113.
- [3] Brasche, G., and B. Walke, "Concepts, services and protocols of the new GSM phase 2+ general packet radio service," *IEEE Communications Magazine*, Aug. 1997, pp. 94-104.
- [4] Postel, J., "Transmission control protocol," *Internet RFC 793*, 1981.
- [5] Jacobson, V., "Congestion avoidance and control," *ACM SIGCOMM Computer Communication Review*, vol. 18, no. 4, pp. 314-329, Aug. 1988.
- [6] Allman, M., V. Paxson, and W. R. Stevens, "TCP congestion control," *RFC 2581*, Apr. 1999.
- [7] Lee, W.C.Y. (1993). *Mobile Communication Design Fundamentals*. 2nd Edition, John Wiley and Sons.
- [8] Inamura, H., G. Montenegro, R. Ludwig, A. Gurtov, and F. Khafizov, "TCP over second (2.5G) and third generation (3G) wireless networks," *Internet draft*, May 2002, URL: <http://www.ietf.org/internet-drafts/draft-ietf-pilc-2.5g3g-08>.
- [9] Floyd, S., and K. Fall, "Promoting the use of end-to-end congestion control in the Internet," *IEEE/ACM Transactions on Networking*, May 1993, URL: <http://www-nrg.ee.lbl.gov/floyd>
- [10] Cáceres, R., and L. Iftode, "Improving the performance of reliable transport protocols in mobile computing environments," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 5, pp. 850-857, June 1995.
- [11] Balakrishnan, H., S. Seshan and R. Katz, "A comparison of mechanisms for improving TCP performance over wireless links," *IEEE/ACM Transactions on Networking*, vol. 5, pp. 756-769, Dec. 1997.
- [12] Balakrishnan, H., S. Seshan, and R. H. Katz, "Improving reliable transport and handoff performance in cellular wireless networks," *ACM/Baltzer Wireless Networks*, vol. 1, no. 4, pp. 469-481, Dec. 1995.
- [13] Balakrishnan, H., S. Seshan, E. Amir, and R. H. Katz, "Improving TCP/IP performance over wireless networks," *Proc. ACM MOBI-COM'95*.
- [14] Brown, K., and S. Singh, "M-TCP: TCP for mobile cellular networks," *Proc. IEEE INFOCOM'96*, 1996.

- [15] Chrungoo, A., V. Gupta, H. Saran, and R. Shorey, "TCP K-SACK: A simple protocol to improve performance over lossy links," *Proc. IEEE GLOBECOM'01*, San Antonio, Texas, USA, Nov. 2001.
- [16] Lin, D., and H. Hung, "TCP fast recovery strategies: Analysis and improvements," *Proc. IEEE INFOCOM'98*, Apr. 1998.
- [17] Casetti, C., M. Gerla, S. Mascolo, M. Y. Sanadidi, and R. Wang, "TCP Westwood: End-to-end congestion control for wired/wireless networks," *Wireless Networks*, vol. , no. 8, pp. 467-479, 2002.
- [18] Parsa, C., and J. J. Garcia-Luna-Aceves, "Improving TCP congestion control over Internets with heterogeneous transmission media," *Proc. IEEE Int. Conference on Network Protocols (ICNP'99)*, Toronto, Canada, Oct. 31-Nov. 3, 1999.
- [19] Varshney, U., "Selective slow start: A simple algorithm for improving TCP performance in wireless ATM environment," *Proc. IEEE MILCOM'97*, pp. 465-469.
- [20] Goff, T., J. Moronski, D. S. Phatak, and V. Gupta, "Freeze-TCP: A true end-to-end TCP enhancement mechanism for mobile environments," *Proc. IEEE INFOCOM'00*.
- [21] Banerjee, D. N., "Improving wireless-wireline TCP interaction," submitted to *IEEE/ACM Transactions on Networking*.
- [22] Sinha, P., N. Venkitaraman, T. Nandagopal, R. Sivakumar, and V. Bharghavan, "WTCP: A reliable transport protocol for wireless wide-area networks," *Proc. ACM MOBICOM'99*, Seattle, Washington, Aug. 1999.
- [23] Ramani, R., and A. Karandikar, "Explicit congestion notification (ECN) in TCP over wireless networks," *Proc. IEEE Int. Conference on Personal Wireless Communications (ICPWC'00)*, pp. 495-499.
- [24] Floyd, S., "TCP and explicit congestion notification," *ACM Computer Communication Review*, vol. 24, no. 5, pp. 10-23, Oct. 1994.
- [25] Ludwig, R., and R. H. Katz, "The Eifel algorithm: Making TCP robust against spurious retransmissions," *ACM Computer Communication Review*, vol. 30, no. 1, Jan. 2000.
- [26] Vaidya, N., Overview of work in mobile-computing (Transparencies), <http://www.cs.tamu.edu/faculty/vaidya/slides.ps>
- [27] Bakre, A., and B. R. Badrinath, "I-TCP: Indirect TCP for mobile hosts," *Proc. 15th IEEE Int. Conf. Distributed Computing Systems (ICDCS)*, pp. 136-143, May 1995.
- [28] Balakrishnan, H., and R. Katz, "Explicit loss notification and wireless web performance," *Proc. IEEE GLOBECOM'98, Internet Mini Conference*, Sydney, Australia.

- [29] Goel, S., and D. Sanghi, "Improving TCP performance over wireless links," *Proc. of IEEE Region Ten Conference on Global Connectivity in Energy, Computer Communication and Control (TENCON'98)*, Dec. 1998.
- [30] Biaz, S., and N. Vaidya, "TCP over wireless networks using multiple acknowledgements," Texas A&M University, Technical Report 97-001, Jan. 1997.
- [31] Wang, K.-Y, and S. K. Tripathi, "Mobile-end transport protocol: An alternative to TCP/IP over wireless links," *Proc. IEEE INFOCOM'98*.
- [32] Ratnam, K., and I. Matta, "WTCP: An efficient mechanism for improving TCP performance over wireless links," *Proc. Third IEEE Symposium on Computers and Communications (ISCC'98)*, Athens, Greece, June 1998.
- [33] Chan, M. C, and R. Ramjee, "TCP/IP performance over 3G wireless links with rate and delay variation," *Proc. ACM MOBICOM'02*, Sept. 2002.
- [34] DeSimone, A., M. C. Chuah, and O. C. Yue, "Throughput performance of transport-layer protocol over wireless LANs," *Proc. IEEE GLOBECOM'93*, Dec. 1993.
- [35] Parsa, C., and J. J. Garcia-Luna-Aceves, "Improving TCP Performance over wireless networks at the link layer," *ACM Mobile Networks and Applications*, Special Issue on Mobile Data Networks: Advanced Technologies and Services, vol. 5, no. 1, 2000, pp. 57-71.
- [36] Vaidya, N., and M. Mehtha, "Delayed duplicate acknowledgements: A TCP-unaware approach to improve performance of TCP over wireless links," Texas A&M University, Technical Report 99-003, Feb. 1999.
- [37] Chiasserini, C.-F., and M. Meo, "Improving TCP over wireless through adaptive link layer setting," *Proc. IEEE GLOBECOM'01*, San Antonio, TX, Nov. 2001.
- [38] Wong, J. W. K., and V. C. M. Leung, "Improving end-to-end performance of TCP using link-layer retransmissions over mobile internetworks," *Proc. IEEE ICC'99*, pp. 324-328.
- [39] Henderson, T.R., E. Sahouria, S. McCanne, and R. Katz, "On improving the fairness of TCP congestion avoidance," *Proc. IEEE GLOBECOM'98*, pp. 539-544.
- [40] Kumar, A., "Comparative performance analysis of versions of TCP in a local network with a lossy link," *IEEE/ACM Transactions on Networking*, vol. 6, no. 4, pp. 485-498, Aug. 1998.

- [41] Lakshman, T.V., and U. Madhow, "The performance of TCP/IP for networks with high bandwidth-delay products and random loss," *IEEE/ACM Transactions on Networking*, vol. 5, no. 3, pp. 336-350, June 1997.
- [42] Fall, K., and S. Floyd, "Simulation-based comparisons of Tahoe, Reno, and Sack TCP," *ACM Computer Communication Review*, Jul. 1996.
- [43] Mathis, M., J. Mahdavi, S. Floyd, and A. Romanow, "TCP selective acknowledgement options," *Internet RFC 2018*, 1996.
- [44] Floyd, S., and T. Henderson, "The New-Reno modification to TCP's fast recovery algorithm," *RFC 2582*, Apr. 1999.
- [45] Mathis, M., J. Mahdavi, S. Floyd, and A. Romanow, "TCP selective acknowledgement options," *RFC 2018*, Apr. 1996.
- [46] Mathis, M., and J. Mahdavi, "Forward acknowledgement: Refining TCP congestion control," *Proc. ACM SIGCOMM'96*, pp. 281-291, 1996.
- [47] Brakmo, L.S., and L. L. Peterson, "TCP Vegas: End to end congestion avoidance on a global Internet," *IEEE Journal on Selected Areas in Communications*, 1995.
- [48] McCanne, S., and S. Floyd, "NS (Network Simulator)," 1995. URL <http://www.isi.edu/nsnam/ns>.
- [49] Chockalingam, A., M. Zorzi, L. B. Milstein, and P. Venkataram, "Performance of a wireless access protocol on correlated Rayleigh fading channels with capture," *IEEE Transactions on Communications*, vol. 46, pp. 644-655, May 1998.
- [50] Chiu, D., and R. Jain, "Analysis of increase and decrease algorithms for congestion avoidance in computer networks," *Computer Networks and ISDN Systems*, vol. 17, pp. 1-14, June 1989.

This page intentionally left blank

Chapter 10

MULTI-SERVICE WIRELESS INTERNET LINK ENHANCEMENTS

George Xylomenos and George C. Polyzos

Mobile Multimedia Laboratory

Department of Informatics

Athens University of Economics and Business

xgeorge@aueb.gr, polyzos@aueb.gr

Abstract The deployment of several real-time multimedia applications over the Internet has motivated a considerable research effort on the provision of multiple services on the Internet. In order to extend this work over wireless links however, we must also take into account the performance limitations of wireless media. We survey various related approaches and conclude that link layer schemes provide a universal and localized solution. Based on simulations of application performance over many link layer schemes we show that different approaches work best for different applications. We present a multi-service link layer architecture which enhances the performance of diverse applications by concurrently supporting multiple link layer schemes. Simulations of multiple applications executing simultaneously show that this approach dramatically improves performance for all of them. We finally consider embedding this approach into a Quality of Service oriented Internet, discussing the traditional best-effort architecture, the Differentiated Services architecture and an advanced dynamic service discovery architecture.

Keywords: Wireless link layer protocols, quality of service, differentiated services.

1. Introduction

The most important Internet related developments in the past few years were arguably the emergence of mechanisms for providing *Quality of Service* (QoS) guarantees to applications and the wide deployment of wireless access networks using technologies such as *Digital Cellular Communications* and *Wireless Local Area Networks* (WLANs). The

QoS effort is largely driven by the desire to migrate applications with real-time delay constraints, such as voice telephony and video conferencing, from the circuit-switched telephony networks to the packet-switched Internet. Many approaches have been proposed for the provision of adequate QoS for such applications, without requiring dramatic changes to the Internet [1, 2]. These schemes provide differentiated treatment to different traffic classes, in terms of packet priorities and drop probabilities, so as to combine the economies of statistical multiplexing with the delay guarantees needed for real-time applications.

Integrating wireless networks into the Internet is relatively simple, due to the physical layer independence of the *Internet Protocol* (IP), which offers a standard interface to higher layers. However, while providing IP services over wireless links is easy, the resulting performance is often disappointing due to the relatively high error rates of wireless links. In addition, while satellite and terrestrial microwave links have long been part of the Internet, higher layer protocols commonly make assumptions about link performance that cannot be met by wireless links, leading to further performance degradation.

This chapter describes an architecture that improves the performance of diverse Internet applications while also extending Internet QoS support over wireless networks. In Sec. 10.2 we outline the problem and review previous work on wireless link enhancements and service differentiation. In Sec. 10.3 we describe the simulation setup that we use throughout the chapter. In Sec. 10.4 we present single-service simulations showing that different applications favor different link enhancement schemes. In Sec. 10.5 we present a *multi-service link layer* architecture that simultaneously enhances the performance of diverse applications by supporting multiple link mechanisms in parallel. In Sec. 10.6 we present simulations showing that with this architecture each application achieves similar gains as when it operates by itself over its preferred link layer mechanism. We then discuss the integration of this approach with various Internet QoS schemes: Sec. 10.7 covers the traditional best-effort service, Sec. 10.8 the Differentiated Services architecture, and Sec. 10.9 an advanced dynamic service discovery architecture.

2. Background and Related Work

IP offers an unreliable packet delivery service between any two Internet hosts. IP packets may be lost, reordered or duplicated. Applications can use the *User Datagram protocol* (UDP) for direct access to this service. Some applications employ UDP assuming that the network is reliable enough for their needs, for example, file sharing via NFS over

wired LANs. Delay sensitive applications may also use UDP in order to employ their own customized loss recovery mechanisms. For example, delay sensitive applications may add redundancy to their data so as to tolerate some losses without (slow) end-to-end retransmissions.

Most applications however prefer complete reliability, hence they employ the *Transmission Control Protocol* (TCP), which offers a reliable byte stream service. TCP breaks the application data stream into segments which are reassembled at the receiver. The receiver returns cumulative *acknowledgments* (ACKs) for segments received in sequence, with duplicate ACKs for out of sequence ones. Since IP may reorder packets, the sender retransmits the first unacknowledged segment only after receiving multiple (usually 3) duplicate ACKs. The sender tracks the round trip delay of the connection, so that if an ACK for a segment does not arrive in time, the segment is retransmitted [3]. Due to the high reliability of wired links, TCP assumes that all losses signify congestion. Therefore, after a loss is detected, TCP reduces its transmission rate, and then increases it slowly so as to gently probe the network [3].

With wireless links, some common assumptions about network reliability are no longer valid. The high error rate of wireless links causes UDP application performance to degrade or even become unacceptable, since it is up to the application to recover from any losses. On the other hand, TCP continuously reduces its transmission rate due to wireless errors to avoid what it (falsely) assumes to be congestion. With multiple wireless links on the path, throughput is reduced in a multiplicative manner [4]. For example, WLANs depict losses of up to 1.5% for Ethernet size frames [5], while cellular systems suffer from losses of 1–2% for their much shorter frames [6]. The effects of these losses are dramatic: a 2% frame loss rate over a WLAN halves TCP throughput [7].

The performance of UDP applications over wireless links has not been studied extensively, not only due to their diversity, but also because they were perceived as oriented towards wired LANs, an assumption challenged by media streaming over the Internet. Considerable work has been devoted to TCP however. Most TCP enhancements try to avoid triggering congestion recovery due to wireless errors. Generic TCP enhancements such as *Eifel* [8] and *Selective Acknowledgments* [9] improve TCP performance by reducing the *number* of redundant TCP retransmissions, without however reducing their (end-to-end) *delay*.

One wireless specific approach is to *split* TCP connections into one connection over the wireless link and another one over the wired part of the path, bridged by an agent [10]. Recovery is only performed locally over the wireless link. This scheme violates the end-to-end semantics of TCP and it is incompatible with IP security which encrypts TCP

headers [11]. Other approaches maintain TCP semantics by *freezing* TCP state whenever persistent errors are detected [12, 13]. As long as errors persist, TCP does not invoke congestion control. Wireless errors and congestion losses are differentiated using either explicit loss notifications [12] or explicit congestion notifications [13]. Thus, performance is not unnecessarily degraded, but error recovery remains end-to-end and the network must provide explicit loss or congestion notifications.

The main alternative to TCP modifications is local error recovery at the link layer. One option is to use the *Radio Link Protocols* (RLPs) of cellular systems which provide error control customized for the underlying link [6, 14]. Since these RLPs only apply a single error control scheme though, they may be inappropriate for some traffic. For example, retransmissions are beneficial to TCP but problematic for real-time traffic. Another problem is that RLPs may interfere with TCP recovery [15], leading to conflicting retransmissions between the link and transport layers. This is avoided by only performing retransmissions when the link layer recovers from losses much faster than the transport layer [16].

One method of jointly optimizing recovery between layers is to exploit transport layer information at the link layer. By *snooping* inside *each* TCP stream at the wireless base station, we can transparently retransmit lost segments when duplicate ACKs arrive, hiding the duplicates from the sender to avoid end-to-end recovery. This approach avoids both control overhead and adverse cross layer interactions [7]. However, it is also incompatible with IP security as it employs TCP header information and it only works in the direction from the wired Internet towards the wireless host due to its reliance on TCP ACKs. In the reverse direction ACKs are returned late and they may signify congestion losses [4].

In order to prevent RLPS from adversely affecting non-TCP traffic, it has been suggested that link layer traffic should be split in two classes, TCP and UDP-based, so as to provide reliable transmission for TCP traffic only. Such a link layer may be either aware or unaware [17, 18] of TCP semantics, as long as it can distinguish TCP from UDP packets. This method does not improve the performance of UDP applications however, which may then fail over the error prone wireless links.

In general, link layer approaches have the advantage over higher layer ones of providing a local solution that does not require any changes to higher layers. This makes them transparent to the rest of the Internet, enables them to recover faster than transport layer solutions and allows them to exploit lower layer information to optimize recovery [16, 18]. Therefore, for the remainder of this chapter we will focus on link layer mechanisms that are able to provide customized error control to each traffic class and the introduction of such mechanisms into the Internet.

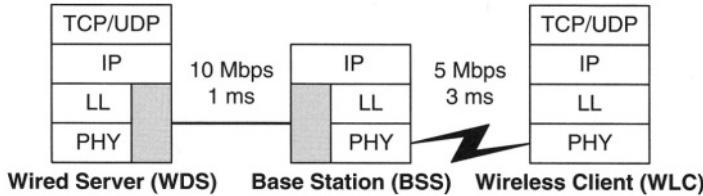


Figure 10.1. Simulation topology.

3. Simulation Setup

To study the interactions between link layer schemes, transport layer protocols and applications, we performed extensive simulations using the ns-2 simulator [19], enhanced with additional error modules, link layer schemes and applications [20]. To compensate for statistical fluctuations, we repeated each experiment 30 times, using the first 30 random seeds embedded in ns-2. The results shown below represent averaged values from all runs, with error bars at plus/minus one standard deviation. We provide elsewhere additional simulation details, as well as results for other network topologies and wireless links [21, 22]. Here we only provide a sample of our results so as to motivate our design.

The simulated network topology is shown in Fig. 10.1. A *Wired Server* (WDS) communicates with a *Wireless Client* (WLC) via a *Base Station* (BSS). The wired link between the WDS and the BSS is a LAN with 10 Mbps of bandwidth and a 1 ms delay. The wireless link between the BSS and the WLC simulates an IEEE 802.11b WLAN with 5 Mbps of bandwidth and a 3 ms delay, using 1000 byte frames. All links are treated as full-duplex pipes for simplicity. To allow comparisons with other studies, the wireless link corrupts bits at exponentially distributed intervals with average durations of 2^{14} , 2^{15} , 2^{16} and 2^{17} bits [7], leading to frame loss rates of 0.8% to 5.9%. We also ran experiments under error-free conditions for reference purposes. The error processes in each wireless link direction are identical but independent. We ignored TCP/UDP/IP headers as they uniformly influence all link layer schemes, but accounted for the *exact* framing overhead required by each link layer scheme.

We tested both TCP and UDP applications, so as to evaluate the suitability of various link layer schemes for diverse applications. For TCP, we simulated file transfers and World Wide Web browsing, using the TCP Reno module of ns-2 with 500 ms granularity timers. For UDP,

we simulated continuous media distribution, a delay sensitive but error tolerant application. In all cases, most data flows from the WDS to the WLC, simulating a client-server interaction with the client on a wireless network. Some data also flow in the reverse direction, for example, TCP ACKs. As a baseline, we simulated a *Raw Link* scheme, which does not perform any error recovery, under both TCP and UDP.

For TCP applications, we tested reliable link layer schemes delivering frames in sequence to higher layers so as to avoid TCP retransmissions. In *Go Back N* the sender buffers outgoing frames and retransmits unacknowledged frames after a timeout. The receiver positively acknowledges frames received in sequence, dropping out of sequence ones. After a timeout the sender retransmits *all* outstanding frames. *Selective Repeat* improves upon this by buffering out of sequence frames at the receiver and returning *negative ACKs* (NACKs) when it detects gaps, thus allowing the sender to retransmit lost frames only. Our Selective Repeat variant allows multiple NACKs per loss [23]. In both schemes, under persistent losses the sender may exhaust its window and stall. Each frame includes sequence and acknowledgment numbers (2 bytes).

To prevent conflicts with TCP retransmissions due to multiple link layer retransmissions [15], the sender in *Karn's RLP* abandons frames not received after 3 retransmissions (for TCP) [6]. Thus, the sender never stalls. This scheme uses only NACK and *keepalive* frames during idle periods, thus frames only include a sequence number (1 byte). Finally, *Berkeley Snoop* is a TCP aware scheme [7], included in the ns-2 distribution. A module at the BSS *snoops* inside TCP segments, buffering data sent to the WLC. If duplicate TCP ACKs indicate a lost packet, this is retransmitted by the BSS and the ACKs are suppressed.

For the UDP-based application simulated, low delay is preferable to full reliability. *Forward Error Correction* (FEC) schemes offer limited recovery by adding redundancy to the transmitted stream, allowing the receiver to recover from losses. The *XOR based FEC* scheme sends data frames unmodified, but every 12 frames (a tradeoff between overhead and delay) a *parity* frame is also transmitted, generated by XOR'ing the preceding data frames, collectively called a *block*. If a *single* data frame is lost from a block, we can recover it by XOR'ing the remaining data frames with the parity frame. A timeout is used to prematurely emit a parity frame when the link becomes idle before the current block has been completed. All frames include a sequence number (1 byte).

The UDP-based application was tested with Selective Repeat, in order to examine the interactions between a fully reliable scheme and a delay sensitive application. We also tested Karn's RLP, but with 1 retransmission per loss to keep delay low. In this scheme, frame losses cause

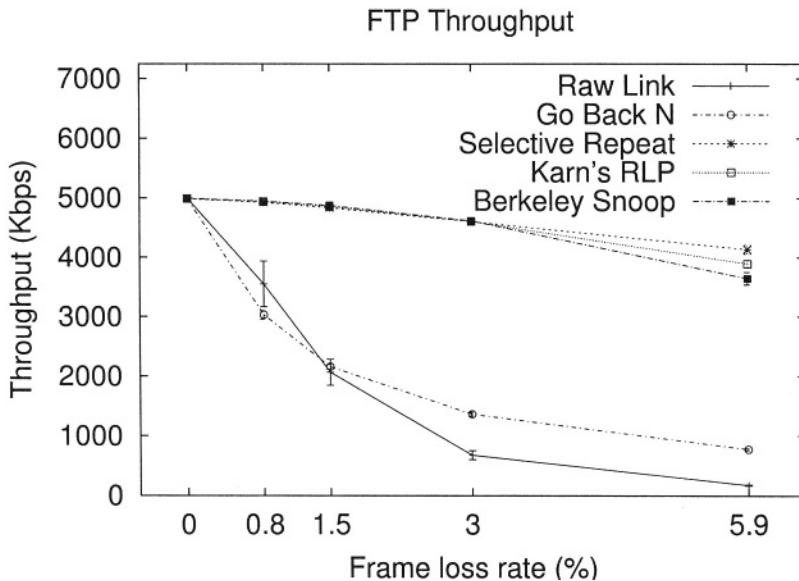


Figure 10.2. Stand-alone file transfer: throughput.

subsequently received frames to wait until the missing one is received or abandoned. This is detrimental to applications using their own resequencing buffers, a common case in continuous media distribution. Our *Out of Sequence* (OOS) variant of Karn's RLP immediately releases all received frames to higher layers so as to avoid such delays. This variant was also tested with 1 retransmission per loss.

4. Single-Service Link Layer Performance

The first TCP-based application tested was file transfer over FTP. We simulated a 100 MByte file transfer from the WDS to the WLC. File transfers are unidirectional, with TCP ACKs travelling in the reverse direction. The ns-2 FTP module sends data as fast as possible, with TCP handling flow and congestion control. As TCP completely controls FTP behavior, performance studies usually rely on FTP transfers in the wired to wireless direction, assumed to be the most common case, so as to characterize TCP performance [7]. We measured application throughput, i.e. the amount of *application* data transferred divided by total time. TCP and link layer retransmissions are *not* included as they signify overhead from the application's viewpoint.

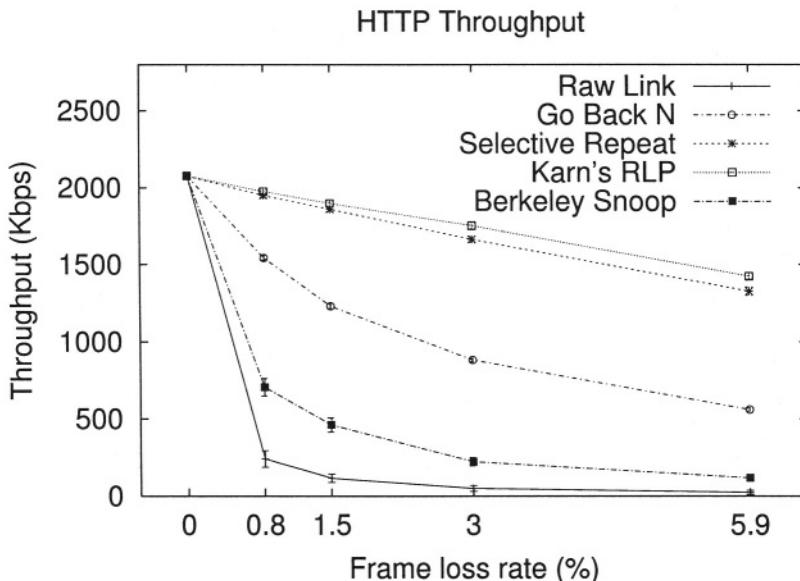


Figure 10.3. Stand-alone WWW browsing: throughput.

Fig. 10.2 shows the FTP throughput achieved by each link layer scheme tested for a range of error rates, averaged over 30 experiments. Most schemes offer large, but similar, performance gains over Raw Link. The exception is Go Back N, due to its naive strategy of retransmitting all outstanding frames after a loss. Due to the high speed of the wireless link, at any given time there are multiple frames in flight over the link. Therefore, retransmitting all outstanding frames after each loss wastes a lot of bandwidth. The other three enhancement schemes are only differentiated at high loss rates, where the most persistent scheme, Selective Repeat, is ahead of Karn's RLP and Berkeley Snoop.

While most studies of wireless TCP performance employ large file transfers, most real applications make many *short* data exchanges, thus TCP rarely reaches the peak throughput suggested by FTP tests. In addition, most applications are either interactive or employ request/reply protocols, thus data flows in *both* directions, and *each* exchange must complete for the application to proceed. Therefore, we also simulated *World Wide Web* (WWW) browsing over HTTP [24], the most popular Internet application. A WWW client accesses *pages* containing text, links and embedded objects, stored on a WWW server. The client-server interaction consists of *transactions*: the client requests a page

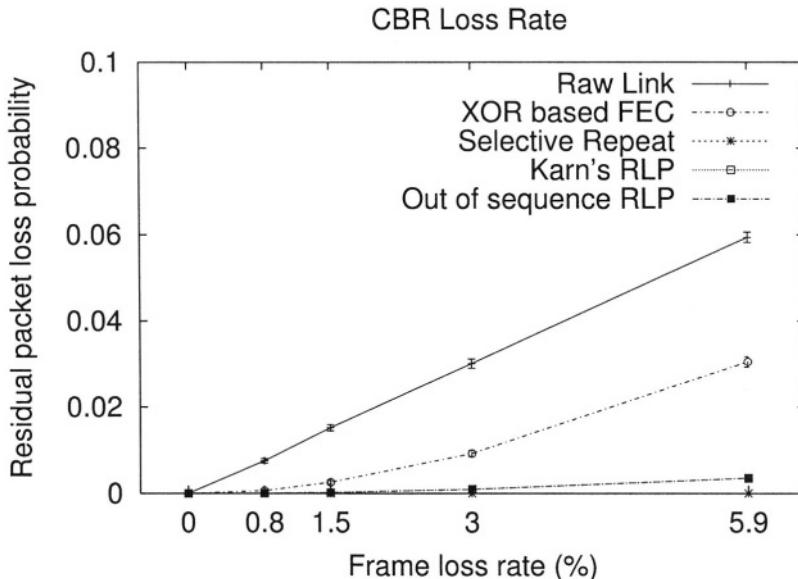


Figure 10.4. Stand-alone continuous media distribution: loss.

from a server, the server returns the page, the client requests all embedded objects, and the server returns them. All transfers are performed over TCP. The ns-2 HTTP module provides empirical distributions for the request, page and embedded object sizes, as well as for the number of objects per page [24]. Only one transaction is in progress at any time and there are no pauses between transactions. WWW browsing was simulated between the WDS and the WLC for 500 s. We measured WWW browsing throughput, i.e. the amount of *application* data transferred from the WDS to the WLC, including both pages and embedded objects, divided by total time.

Fig. 10.3 shows the WWW browsing throughput achieved by each link layer scheme tested. Since the short data transfers of WWW browsing rarely reach high speeds, the naive retransmission strategy of Go Back N does not cause a very dramatic performance deterioration. Karn's RLP performs slightly better than Selective Repeat, again due to the small transfers, since its keepalive feature allows it to recover fast from losses when the link becomes idle, unlike Selective Repeat which has to wait for a timeout to detect losses. The major difference is in the performance of Berkeley Snoop, which provides only minor improvements over Raw Link, as it does not retransmit in the WLC to BSS direction. Thus, even

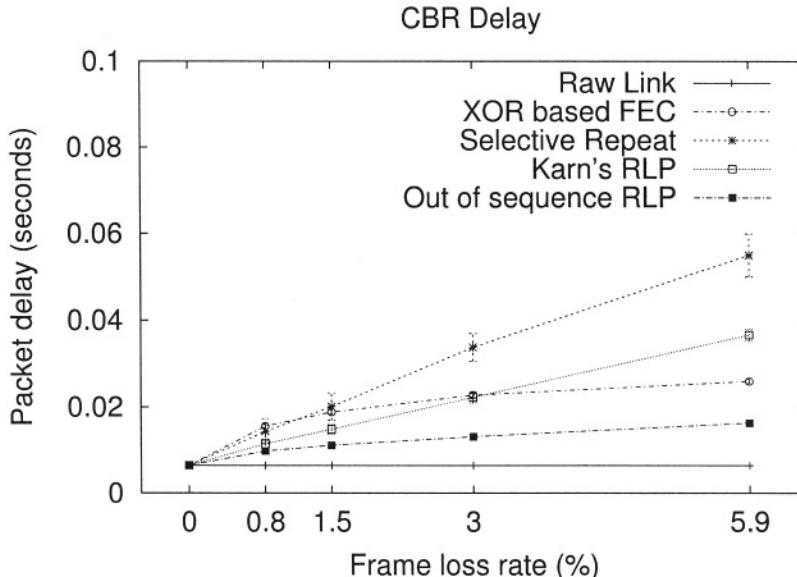


Figure 10.5. Stand-alone continuous media distribution: delay.

if most traffic flows in the server to client direction, in an interactive application the traffic in the client to server direction is equally critical for overall application performance.

Turning now to UDP, while applications using UDP due to its simplicity would work well with all of the above schemes, delay sensitive applications would face serious problems with a fully reliable retransmission scheme. We thus tested UDP performance using real-time continuous media distribution, i.e. a lecture where a speaker at the WDS sends audio and video to an audience including the WLC. We used a speech model where the speaker alternates between *talking* and *silent* states with exponential durations, averaging 1 s and 1.35 s, respectively [25]. Media are only transmitted in the talking state. When talking, the speaker transmits packets isochronously at a *Constant Bit Rate* (CBR) of 1 Mbps. We simulated continuous media distribution for 500 s. We assumed that the application uses FEC to tolerate some loss without retransmissions and that received packets are buffered until a *playback point* determined by human perception. To characterize performance, we measured the *residual* loss rate at the receiver after link layer recovery. Since packets missing the playback point are dropped, loss rate must be coupled with a delay metric covering *most* packets. We used as

a metric mean packet delay plus twice its standard deviation, so as to incorporate the effects of variable delays.

Fig. 10.4 shows residual loss with each scheme. Both RLP schemes dramatically reduce losses, even with a single retransmission per loss. In contrast, XOR based FEC depicts gains that do not justify its overhead. For example, by introducing 8.3% of overhead, a native loss rate of 3% is reduced to 1%. Selective Repeat offers full reliability, but at the cost of very high delays. Fig. 10.5 shows the delay metrics of each scheme with continuous media distribution. The Raw Link curve is flat since no recovery takes place. Selective Repeat exhibits the highest delay due to its full recovery. While both RLP variants perform similarly at low error rates, as link conditions deteriorate only OOS RLP manages to keep delay low, since in sequence delivery causes many packets to be delayed after each loss. Interestingly, XOR based FEC is slower than OOS RLP. The reason is that while OOS RLP requires a NAK and a retransmission to recover from a loss, the FEC scheme requires half a block of frames to be received, on average, in order to recover a lost frame. Since this wireless link has low delay, high bandwidth and uses large frames, retransmission turns out to be faster than this FEC scheme.

5. Multi-Service Link Layer Architecture

The results presented above show that error recovery at the link layer can considerably enhance Internet application performance over wireless links. They also show however that different schemes work best for the TCP and UDP applications tested. This means that a single link layer scheme *cannot* optimize the performance of all TCP and UDP applications. We have therefore developed a *multi-service link layer* architecture, supporting multiple link enhancement services in parallel over a single physical link [26]. Each service fits the needs of an application class, such as TCP-based or real-time UDP-based. Since the adjacent protocol layers expect the link layer to have single entry and exit points, our architecture provides sublayers to assign incoming packets to services and to multiplex outgoing packets from all services into a single stream. These sublayers keep services unaware of the fact that they are operating within a multi-service context.

Fig. 10.6 outlines our architecture, showing data flow in one direction. Incoming packets are classified and passed to the most appropriate service, based on their application class. A simple classifier may use the protocol field of the IP header to distinguish between TCP and UDP, and the port field of the TCP/UDP header to determine the application in use. When *Differentiated Services* (DS) are used, the classifier may

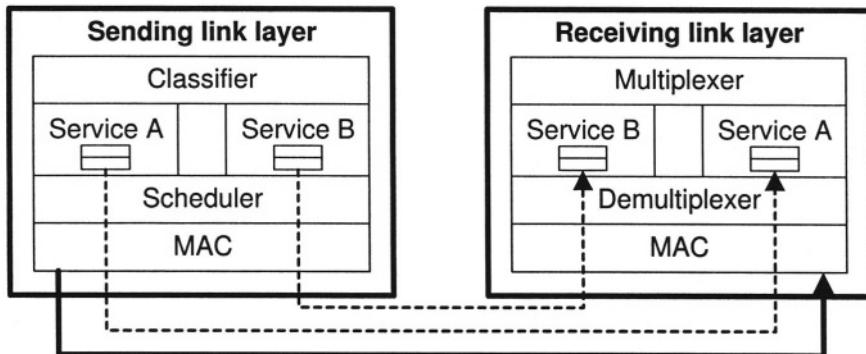


Figure 10.6. Multi-service link layer architecture.

use the DS field of the IP header [1], which remains visible even with IP security [11]. When *Integrated Services* [2] are used, packets can be classified based on flow state maintained by higher layers. Packets from unknown applications are mapped to the default, best-effort, service.

Each service may employ retransmissions, FEC, or any other mechanism desired, keeping its private buffers, counters and timers. Outgoing frames are passed to a scheduler which tags each frame with a service number and eventually delivers it to the MAC sublayer for transmission. At the receiver, frames are passed by the MAC sublayer to the demultiplexer which checks their tags and delivers them to the proper service. Services may eventually release data to the multiplexer, which passes them to the network layer. Therefore, the peer link layer services communicate over *virtual links*.

As each service may arbitrarily inflate its data stream with error recovery overhead, if we simply transmit all frames in a FIFO manner we will penalize the services introducing less overhead. Therefore, we have introduced a frame scheduler to ensure that the link is *impartially* shared between services. Impartiality means that each service should receive the same amount of bandwidth as in a single-service link layer. That is, a service performing no error recovery will get the same bandwidth for data in both cases. In contrast, an error recovery service will get the same *total* bandwidth, but it will have to divide it between its data and its error recovery overhead.

In order to achieve impartiality, we chose a *Self Clocked Fair Queueing* (SCFQ) scheduler [27] which efficiently, but strictly, enforces the desired bandwidth allocations for each service. When some services are

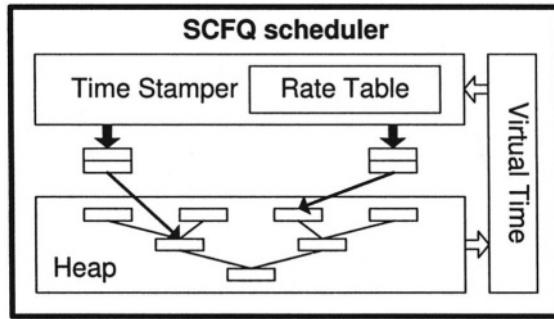


Figure 10.7. Self Clocked Fair Queueing frame scheduler.

idle, their bandwidth is shared among the rest in proportion to their allocations. Fig. 10.7 gives an outline of the scheduler. The *rate table* holds the fraction of link bandwidth allocated to each service. Frames awaiting transmission are buffered in a separate FIFO queue per service. A virtual time variable is maintained, equal to the *time stamp* of the last packet transmitted. To determine the time stamp of an incoming packet, we divide its size by its service rate, and add the result to the time stamp of the preceding frame in its queue. If this queue is empty, we add the result to the current virtual time. When the link is idle, the frame with the *lowest* virtual time is dequeued, its time stamp becomes the system's virtual time, and the frame is transmitted.

Since the frames entering each of the FIFO queues have increasing time stamps, we only need to check the head of each queue to determine which frame has the lowest virtual time. The scheduler keeps all non-empty queues in a heap, sorted by the time stamp of their first frame. The heap is sorted when a frame is dequeued for transmission, based on the new head of the corresponding queue. Empty queues are removed from the heap and are re-inserted when they receive a new frame, which also causes the heap to be sorted. Thus, each frame requires $O(\log_2 n)$ operations (for n services) to sort the heap when the frame leaves (and, possibly, when it enters) the scheduler.

This multi-service link layer architecture can be locally deployed over isolated wireless links, transparently to the rest of the Internet. Additional services can be provided by inserting new modules and extending the mappings of the classifier. Services may be optimized for the underlying link, freely selecting the most appropriate mechanisms. The scheduler ensures that these services fairly share the link, despite their variable overheads, thus keeping services unaware of each other.

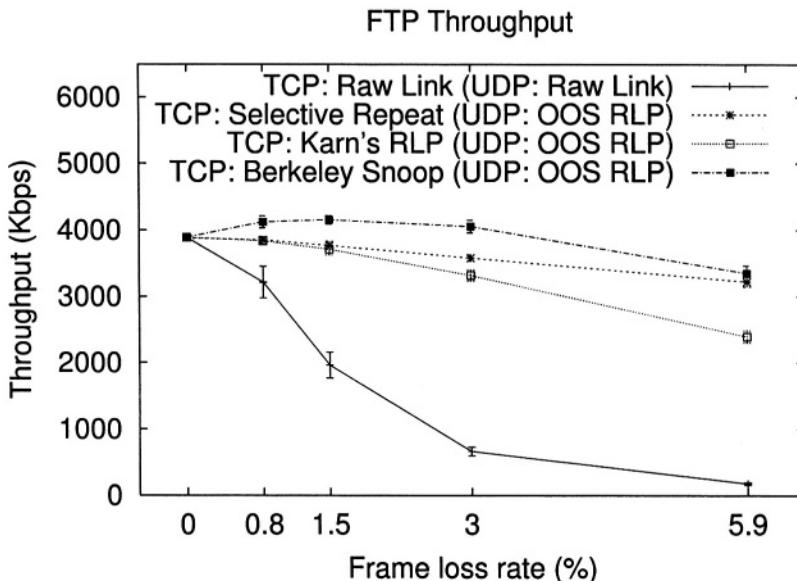


Figure 10.8. All applications: file transfer throughput.

6. Multi-Service Link Layer Performance

To verify whether our architecture can indeed simultaneously enhance the performance of diverse applications, we repeated the simulations of Sec. 10.4 with all applications, i.e. file transfer, WWW browsing and continuous media distribution, executing in parallel but over different link layer services. The TCP servers and UDP sender were at the WDS and the TCP clients and UDP receiver were at the WLC. All applications started together and the simulation ended when the 100 Mbyte file transfer completed. Our multi-service link layer module for ns-2 provided one TCP and one UDP service, using the link layer schemes that performed best in single application experiments. Both TCP applications used the *same*, link layer service. The classifier assigned packets to services based on their DS fields [1] which were set by the applications.

The multi-service link layer module used the SCFQ scheduler described. The rate table was set statically so that continuous media distribution was guaranteed its *peak* bandwidth. Although the scheduler allocated 1 Mbps to the UDP service, the average bandwidth available to the TCP service was 4.575 Mbps, since the UDP application was not constantly active. File transfer is generally more aggressive in terms of

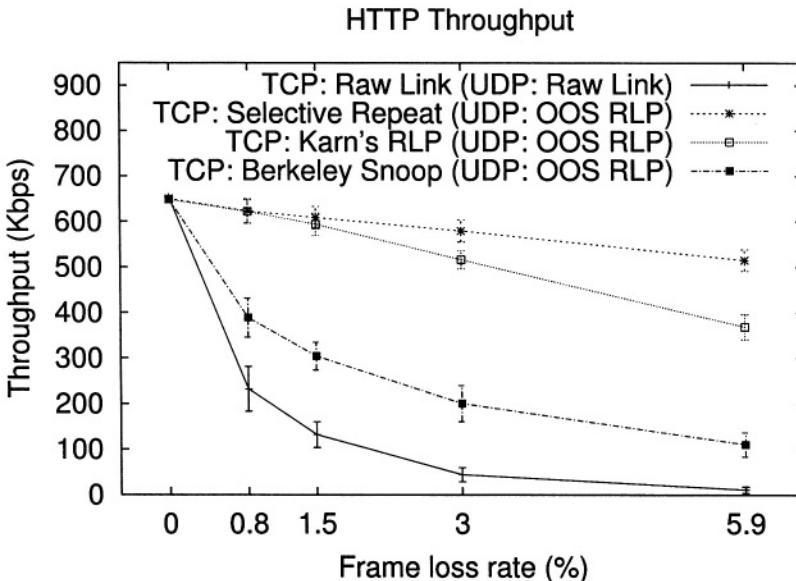


Figure 10.9. All applications: WWW browsing throughput.

competing for bandwidth than WWW browsing, as it performs a bulk transfer. This is handled by existing TCP mechanisms however.

Fig. 10.8 presents file transfer throughput results, with each curve showing the service used for TCP (in parentheses, the service used for UDP). These throughput curves are generally similar to those of single application experiments, but with lower average throughput due to contention with other applications. Selective Repeat performs better than Karn's RLP, especially at higher loss rates, due to its higher persistence. Berkeley Snoop performs best, with its throughput initially increasing with higher loss rates. This is due to a corresponding degradation in WWW browsing throughput with this scheme, as shown below, which leaves more bandwidth available for file transfer.

Results for WWW browsing throughput, shown in Fig 10.9, are also similar to those of single application experiments. The performance of Selective Repeat and Karn's RLP is quite similar to the file transfer throughput results, with very large performance gains compared to Raw Link. Selective Repeat performs better than Karn's RLP as, due to its persistence, it competes for bandwidth more aggressively. Berkeley Snoop fails to offer significant gains due to its topological limitations, exactly as in single application experiments. This drop in WWW browsing

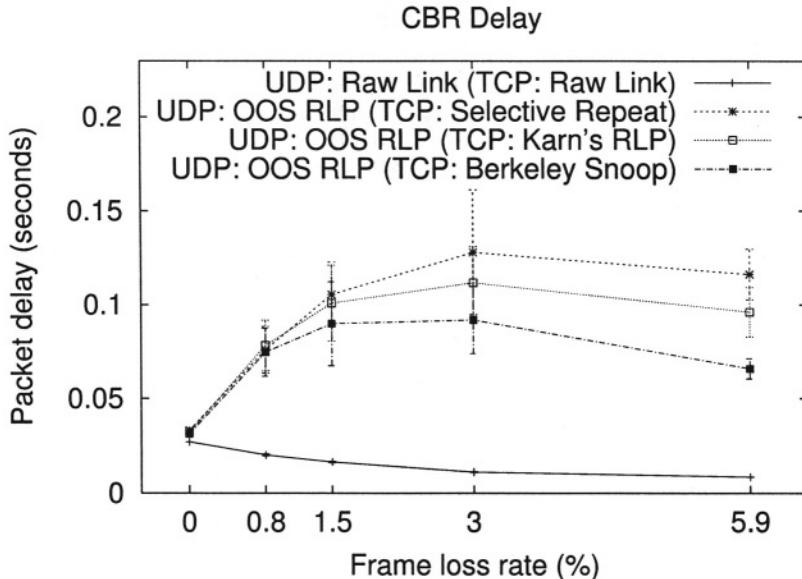


Figure 10.10. All applications: continuous media distribution delay.

throughput with Berkeley Snoop explains why file transfer throughput initially increases with higher loss rates.

Residual losses for continuous media distribution were the same as in single application experiments, since error recovery is not influenced by contention. The delay metrics, given in Fig. 10.10, show for each curve the service used for UDP (in parentheses, the service used for TCP). Delay was inflated, an unavoidable effect with our non-preemptive work-conserving scheduler. The scheduler managed however to keep delay at reasonable levels, especially considering that the additional delay over Raw Link is partly due to the OOS RLP scheme itself. The reduced delays at higher native loss rates were due to the deteriorating TCP performance which reduced contention. Karn's RLP provided a better balance between TCP throughput and UDP delay than the more persistent Selective Repeat. Berkeley Snoop caused the lowest additional delays, at the cost of reduced aggregate TCP performance.

7. Best-effort Service Interface

The network layer of the Internet provides a single, best-effort, packet delivery service. Higher layer protocols can extend this service so as to satisfy additional application requirements. Our simulations show how-

ever that multiple link layer services are needed in order to enhance Internet application performance over wireless links. Since higher layer protocols are not aware of multi-service links, they cannot map their requirements to available services. To retain compatibility with existing infrastructure, our architecture must perform this mapping transparently. This implies that performance should be at least as good as with a single service, that is, applications should not be mapped to inappropriate services and enhancing the performance of one application should not hurt the performance of other applications. These requirements allow services to be introduced gradually to selectively enhance the performance of some applications.

Since our link layer architecture emulates a single service, the only data that can be used for service selection are the contents of incoming IP packets. We can match application requirements to services using a heuristic classifier which recognizes applications based on IP, TCP and UDP packet header fields. Heuristic packet classification starts with the protocol field of the IP header, indicating TCP, UDP, or another protocol. TCP applications can be all mapped to a single service. For UDP applications however, decisions must be made on a per application basis. Known applications can be detected by examining the source and destination port fields of the UDP header. Another field that may be used is the *Type of Service* (TOS) field of the IPv4 header. Originally, this field was intended to indicate application preferences and packet priorities. This field, however, is rarely used and it has been redefined to support Differentiated Services. This also applies to the *Traffic Class* field of the IPv6 header.

Fig. 10.11 shows data flow in such a classifier for IPv4 packets. Headers are masked to isolate the fields used for classification. These fields pass through a hashing function that produces an index to a lookup table, whose entries point at the available services. Unrecognized applications are mapped to the default (native) link service which offers the same performance as a single service link layer. The header mask, hashing function and lookup table are provided by an external entity, i.e. an administrator, which is aware of application requirements, header fields and link services.

Higher layers expect the link layer to allocate bandwidth as if a single service was offered, therefore link services should respect this (implicit) allocation. As packets are assigned to services, the classifier measures their size to deduce the share of the link allocated to each service. Over regular time intervals, the classifier divides the amount of data assigned to each service i , by the total amount of data seen, to get a fraction r_i , where $\sum_{i=1}^n r_i = 1$, for n services. These fractions, or normalized

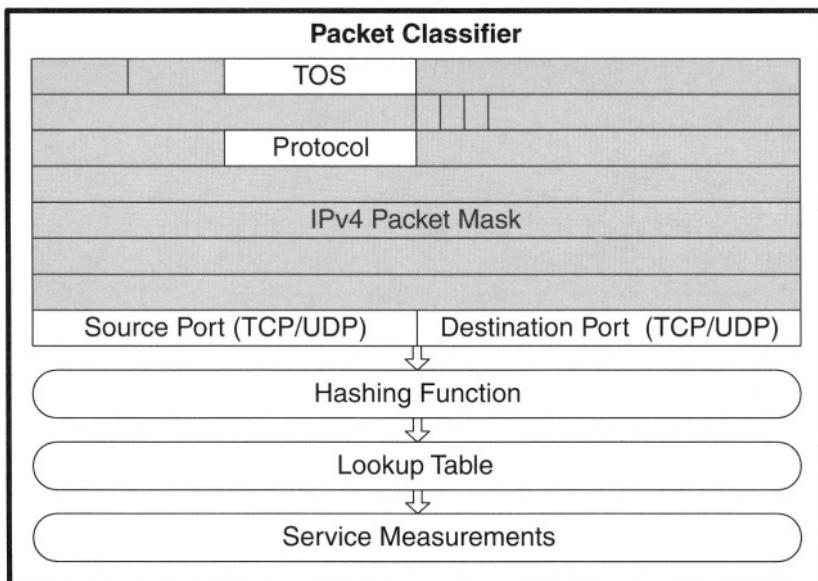


Figure 10.11. Heuristic packet classifier.

service rates, are entered into the rate table. Applications mapped to the default service are allocated the same bandwidth as with a single service. The remaining services trade-off throughput for error recovery.

One drawback of heuristic classifiers is the effort required to construct and maintain them. Applications must be manually matched to services over each type of wireless link whenever new applications or services are added. Many applications will be unrecognized, because they do not use well-known ports. Although some higher layer QoS schemes combine the transport protocol and the source/destination port/address fields to classify packets [28], maintaining state for all these combinations requires end-to-end signaling. Another important problem is that IP security mechanisms encrypt the port fields of TCP and UDP headers and put the identifier of the IP security protocol in the protocol field [11], thus hindering heuristic classification.

8. Differentiated Services Interface

The best-effort service provided by IP is becoming inadequate as real-time multimedia applications are being deployed over the Internet. To support these applications, some type of performance guarantees must be introduced into the Internet. The main issue for Internet QoS is gen-

erally considered to be congestion control, since applications may not be able to get their required throughput due to contention for bandwidth and their end-to-end delay may increase due to queueing delays. One scheme for Internet QoS provision is the *Integrated Services* architecture [2], which has been criticized in two ways. First, it must be widely deployed over the Internet to be useful, since its guarantees rely on actions at every router on a path. Second, it mandates resource reservations on a per flow basis, where a flow is a data stream between two user processes. Since a huge number of flows exists, the scalability of this scheme is questionable.

An alternative scheme that attempts to avoid these limitations is the *Differentiated Services* architecture [1], in which flows are aggregated into a few classes, either when entering the network or when crossing network domains. At these points flows may be rate limited, shaped or marked to conform to specific traffic profiles. For neighboring domains, traffic profiles represent large traffic aggregates, while at network entry points they represent user requirements. Within a domain, routers only need to select a *Per-Hop Behavior* (PHB) for each packet, based on its class, denoted by the 8-bit Differentiated Services (DS) field of the IP header, which subsumes the IPv4 TOS and the IPv6 Traffic Class fields.

The services provided by this architecture are meant to provide generic QoS levels, not application specific guarantees. For example, the expedited forwarding PHB provides guaranteed bandwidth at each router, for traffic that was rate limited when entering the network or the domain so as not to exceed this bandwidth. This PHB provides low delay and loss by eliminating congestion for this restricted traffic class. Only network entry points are aware of both application requirements and PHB semantics so as to perform flow aggregation. Similarly, only domain entry points are aware of the semantics of PHBs available in their neighboring domains so as to perform appropriate translations.

Differentiated Services and multi-service link layers solve orthogonal but complementary problems. Differentiated Services are concerned with congestion and its impact on throughput, delay and loss. The services offered are link independent and are supported by IP level mechanisms. Multi-service link layers are concerned with recovery from link errors, customized to each type of application. The services offered are link dependent and local, as the frame scheduler only protects services from each other. By only providing Differentiated Services over wireless links we can offer a nominal IP level QoS, but the actual performance will be limited by link losses. Multi-service link layers provide adequate recovery to fully utilize wireless links, but they need higher layer guidance to allocate link bandwidth.

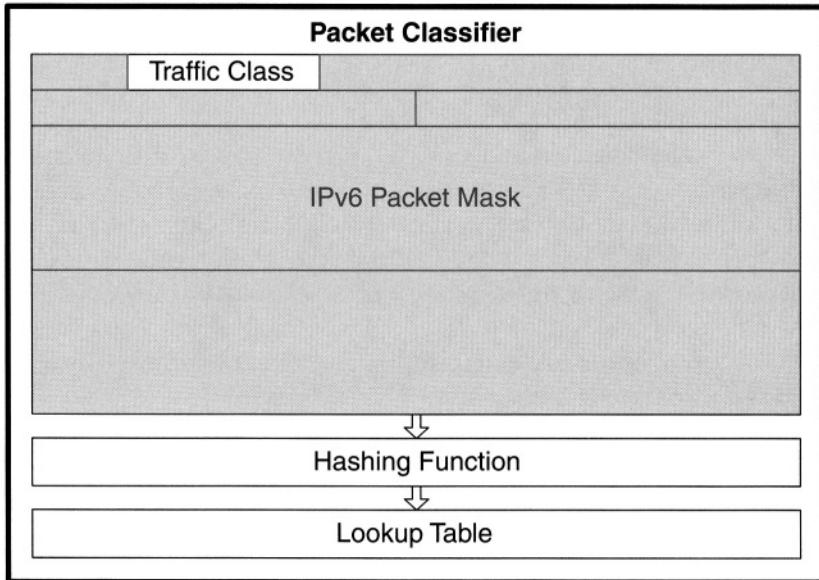


Figure 10.12. Differentiated Services packet classifier.

The two architectures can be implemented so as to complement each other. Differentiated Services provide congestion control, while multi-service link layers add application dependent error control. They both offer few services (PHBs or link mechanisms) for traffic classes with common requirements. They can be combined by extending the DS field to also specify the error requirements of each traffic class. For example, a DS traffic class could be subdivided into two subclasses with different error recovery requirements by using one more bit of the DS field. Applications could indicate their requirements when injecting their traffic into the network, with boundary routers translating them to local equivalents. The result is a simplified classifier, shown in Fig. 10.12 for IPv6 packets. The DS field is isolated via a header mask, and then a hashing function and a lookup table map it to an appropriate service.

This classifier does not rely on multiple header fields and complex rules to determine application requirements. More importantly, the DS field is not obscured by IP security mechanisms. Since the Differentiated Services module performs scheduling, traffic entering the multi-service link layer already obeys the required bandwidth allocations. Therefore, the subclasses of different DS classes can share the same link service without introducing congestion. The service rates can be set in two

Table 10.1. Service characterization metrics.

| Name | Symbol | Definition |
|-------------------|-------------------------|---|
| Goodput | g_i | $\frac{\text{higher layer data transmitted}}{\text{link layer data transmitted}}$ |
| Loss | l_i | $\frac{\text{higher layer data lost}}{\text{higher layer data transmitted}}$ |
| Delay | d_i | Average packet delivery delay |
| Effective Goodput | $e_i = g_i * (1 - l_i)$ | $\frac{\text{higher layer data received}}{\text{link layer data transmitted}}$ |

ways. If the subclasses of each DS traffic class have separate bandwidth allocations at the IP level, then the bandwidths for all subclasses mapped to the same link service are added to set its service rate. Otherwise, a measurement module can be inserted after the lookup table, as in Fig. 10.11, to determine service rates as explained in Sec. 10.7.

9. Advanced Quality of Service Interface

The performance of the services available at each wireless link may vary widely, depending both on the underlying link and on transient conditions. If we could provide a characterization of the end-to-end performance of a network path, applications could verify if a given service was suitable for their needs [29]. If these characterizations were updated whenever path characteristics changed significantly, they would also enable adaptive protocols and applications to modify their policies accordingly. To describe the services offered over diverse network paths however, we must dynamically discover what is provided at each multi-service wireless link on the path. This can be achieved if each service dynamically characterizes its performance with a set of standardized metrics. Dynamic characterization means that performance may be evaluated as often as needed, while metric standardization means that higher layers will be able to assess arbitrary services without any knowledge of the mechanisms employed. This would enable end-to-end QoS modules to compose end-to-end path metrics from local link metrics.

To support such services, we have defined three link independent metrics, reflecting some common trade-offs of error recovery schemes:

- *Goodput* (g_i , for service i) is the ratio of higher layer data transmitted during the measurement interval to link layer data transmitted, including all overhead. Note that the amount of higher layer data transmitted may differ from the amount received due to residual losses. Goodput can be calculated at the sender without any receiver feedback.
- *Loss* (l_i) is the ratio of higher layer data lost to higher layer data transmitted (lost plus received). Loss is calculated by the receiver based on the sequence of data released to higher layers. Residual loss can be greater than zero for limited recovery schemes.
- *Delay* (d_i) is the one way average delay for higher layer packets using this service. Delay can be estimated at the receiver based on knowledge of the implemented recovery scheme and wireless link characteristics. For example, retransmission schemes could add one round trip delay for each actual retransmission to their one way delay estimate.

The delay metric only reflects error recovery delay, so it should be added to IP level queueing delay to give the total delay for a node. Delay sensitive traffic subclasses should choose the lowest delay service whose residual loss falls within their tolerance limits. Goodput can be combined with loss to get *Effective Goodput* (e_i), defined as $e_i = g_i * (1 - l_i)$, or, the ratio of higher layer data received to link layer data transmitted. Essentially, e_i shows how much of the bandwidth allocated to a service is used by data actually received, after subtracting error recovery overhead and residual losses. If the link bandwidth is B and service i is allocated a normalized service rate r_i in the scheduler, its throughput is $B * r_i * e_i$. This expression may be used to estimate the throughput for each service given a set of service rates, or to calculate the service rate needed to achieve a target throughput. All metrics are summarized in Table 10.1.

These metrics may be used at multiple layers to serve different needs, as shown in Fig. 10.13. The physical layer provides hardware information, such as the fixed one-way delay, that may be used by link layer services to provide their link independent metrics. At the network layer, scheduling mechanisms such as *Class Based Queueing* (CBQ) may use those metrics to set bandwidth allocations for each service. End-to-end QoS schemes may use the *Resource ReSerVation Protocol* (RSVP) to gather information about node services so as to estimate end-to-end path characteristics. These can in turn be used by transport protocols

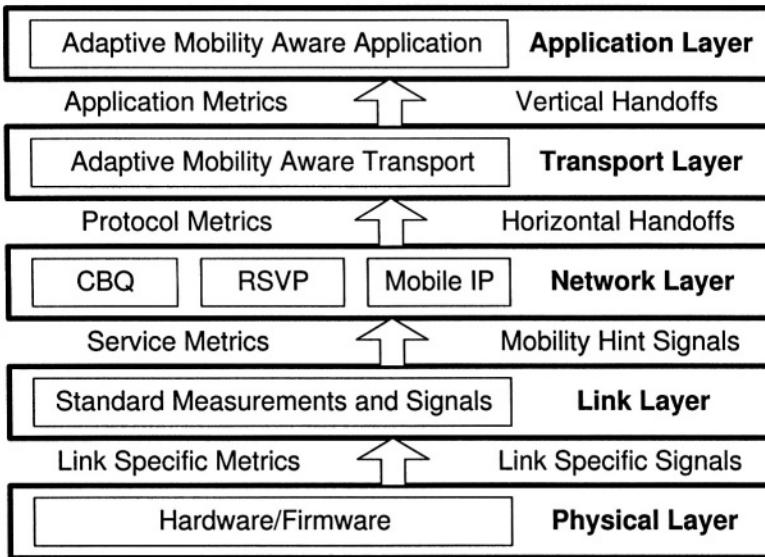


Figure 10.13. Propagation of service measurement and mobility feedback.

and applications to adapt their operation to prevailing conditions. For example, video conferencing applications may select encoding schemes appropriate to the residual loss characteristics of the path.

To assist higher layers in dealing with mobility, we can extend this interface to provide *mobility hints* to interested parties via upcalls, as shown in Fig. 10.13. The link layer can combine hardware signals with its own state to detect events like connections and disconnections. These link independent upcalls can be used by IP mobility extensions to allow fast detection of handoffs, instead of relying on periodic network layer probes [30]. Higher layers may be notified by the network layer via further upcalls of horizontal and vertical handoffs, i.e. handoffs between radio cells employing the same technology and handoffs between radio cells employing different technologies, respectively. TCP may be notified of pending horizontal handoffs to temporarily freeze its timers and avoid timeouts during disconnection. A video conferencing application may be notified of vertical handoffs so as to change the encoding scheme used to a higher or lower resolution one, depending on available bandwidth.

This interface fits the requirements of *One-Pass With Advertising* (OPWA) [29] resource reservation mechanisms. One-pass resource reservation schemes cannot specify a desired service in advance as they do not know what is available on a path [28], thus the resources reserved may

prove to be inadequate. Two-pass resource reservation schemes specify a service in advance but make very restrictive reservations on the first pass, relaxing them on a second pass. Such reservations may fail due to tight restrictions on the first pass. In an OPWA scheme, an advertising pass is first made to discover the services available on the path, and then a reservation pass reserves the resources needed. Our interface allows OPWA schemes to discover the restrictions imposed by wireless links. After that information is gathered, applications choose a service and the reservation pass sets up appropriate state to provide it. Mobility hints notify higher layers that they should revise path characterizations after handoffs, so as to support mobility aware applications.

10. Summary

While extending the Internet over wireless links is straightforward, application performance over wireless links using the Internet protocols is often disappointing due to wireless impairments. These problems considerably complicate the provision of Quality of Service guarantees over wireless Internet links. In order to investigate application performance over wireless links, we have simulated a number of link layer error control schemes. Our results show that different applications favor different approaches, so we have developed a link layer architecture that supports multiple services simultaneously over a single link, allowing the most appropriate link service to be used by each traffic class.

Our approach is easy to optimize for each underlying wireless link and efficient to operate. It can be locally deployed, transparently to the rest of the Internet, and it is easy to extend to address future requirements. It can be incorporated into the Internet QoS architecture in three stages. First, in the current best-effort only Internet, heuristic classifiers can be employed to select the services provided over individual multi-service links, so as to enhance the performance of recognized applications. Second, in the context of the Differentiated Services architecture, which focuses on end-to-end congestion control, multi-service link layers can provide application dependent error control, respecting higher layer scheduling decisions despite the introduction of error recovery overhead. Finally, multi-service link layers can support an advanced QoS application interface, offering dynamic end-to-end service discovery.

References

- [1] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, “An architecture for Differentiated Services,” RFC 2475, December 1998.
- [2] D. D. Clark, S. Shenker, and L. Zhang, “Supporting real-time applications in an integrated services packet network: architecture and mechanism,” in *Proc. of the ACM SIGCOMM ’96*, August 1996, pp. 243–254.
- [3] W. Stevens, “TCP slow start, congestion avoidance, fast retransmit, and fast recovery algorithms,” RFC 2001, January 1997.
- [4] G. Xylomenos and G. C. Polyzos, “Internet protocol performance over networks with wireless links,” *IEEE Network*, vol. 13, no. 5, pp. 55–63, July/August 1999.
- [5] G. T. Nguyen, R. H. Katz, B. Noble, and M. Satyanarayanan, “A trace-based approach for modeling wireless channel behavior,” in *Proc. of the Winter Simulation Conference*, 1996.
- [6] P. Karn, “The Qualcomm CDMA digital cellular system,” in *Proc. of the USENIX Mobile and Location-Independent Computing Symposium*, 1993, pp. 35–39.
- [7] H. Balakrishnan, V. N. Padmanabhan, S. Seshan, and R. H. Katz, “A comparison of mechanisms for improving TCP performance over wireless links,” in *Proc. of the ACM SIGCOMM ’96*, 1996, pp. 256–267.
- [8] R. Ludwig and R. H. Katz, “The Eifel algorithm: making TCP robust against spurious retransmissions,” *Computer Communications Review*, vol. 30, no. 1, pp. 30–36, January 2000.
- [9] M. Mathis, J. Mahdavi, S. Floyd, and A. Romanow, “TCP selective acknowledgment options,” RFC 2018, October 1996.
- [10] B. R. Badrinath, A. Bakre, T. Imielinski, and R. Marantz, “Handling mobile clients: A case for indirect interaction,” in *Proc. of the 4th Workshop on Workstation Operating Systems*, 1993, pp. 91–97.
- [11] S. Kent and R. Atkinson, “IP encapsulating security payload (ESP),” RFC 2406, November 1998.

- [12] T. Goff, J. Moronski, D.S. Phatak, and V. Gupta, "Freeze-TCP: A true end-to-end TCP enhancement mechanism for mobile environments," in *Proc. of the IEEE INFOCOM '00*, 2000, pp. 1537–1545.
- [13] J. Liu and S. Singh, "ATCP: TCP for mobile ad hoc networks," *IEEE Journal on Selected Areas in Communications*, vol. 19, no. 7, pp. 1300–1315, July 2001.
- [14] S. Nanda, R. Eljak, and B. T. Doshi, "A retransmission scheme for circuit-mode data on wireless links," *IEEE Journal on Selected Areas in Communications*, vol. 12, no. 8, pp. 1338–1352, October 1994.
- [15] A. DeSimone, M. C. Chuah, and O.C. Yue, "Throughput performance of transport-layer protocols over wireless LANs," in *Proc. of the IEEE GLOBECOM '93*, 1993, pp. 542–549.
- [16] Q. Pang, A. Bigloo, V.C.M. Leung, and C. Scholefield, "Performance evaluation of retrnasmision mechanisms in GPRS networks," in *Proc. of the IEEE WCNC '00*, 2000, pp. 1182–1186.
- [17] R. Ludwig and B. Rathonyi, "Link layer enhancements for TCP/IP over GSM," in *Proc. of the IEEE INFOCOM '99*, 1999, pp. 415–422.
- [18] C. Parsa and J.J. Garcia-Luna-Aceves, "Improving TCP performance over wireless networks at the link layer," *Mobile Networks and Applications*, vol. 5, no. 1, pp. 57–71, 2000.
- [19] The VINT Project, "UCB/LBNL/VINT Network Simulator - ns (version 2)," Available at <http://www.isi.edu/nsnam>.
- [20] G. Xylomenos, "Multi-service link layer extensions for ns-2," Available at <http://www.mm.aueb.gr/~xgeorge/>.
- [21] G. Xylomenos and G. C. Polyzos, "Multi-service link layer enhancements for the wireless Internet," in *Proc. of the IEEE Symposium on Computers and Communications '03*, 2003, pp. 1147–1152.
- [22] G. Xylomenos and G. C. Polyzos, "Wireless link layer enhancements for TCP and UDP applications," in *Proc. of the International Parallel and Distributed Processing Symposium '03*, 2003, pp. 225–232.
- [23] P. T. Brady, "Evaluation of multireject, selective reject, and other protocol enhancements," *IEEE Transactions on Communications*, vol. 35, no. 6, pp. 659–666, June 1987.

- [24] B. A. Mah, “An empirical model of HTTP network traffic,” in *Proc. of the IEEE INFOCOM '97*, 1997, pp. 592–600.
- [25] S. Nanda, D. J. Goodman, and U. Timor, “Performance of PRMA: a packet voice protocol for cellular systems,” *IEEE Transactions on Vehicular Technology*, vol. 40, no. 3, pp. 584–598, August 1991.
- [26] G. Xylomenos and G. C. Polyzos, “Quality of service support over multi-service wireless Internet links,” *Computer Networks*, vol. 37, no. 5, pp. 601–615, 2001.
- [27] S. Golestani, “A self-coded fair queueing scheme for broadband applications,” in *Proc. of the IEEE INFOCOM '94*, 1994, pp. 636–646.
- [28] L. Zhang, S. Deering, D. Estrin, S. Shenker, and D. Zappala, “RSVP: A new resource reservation protocol,” *IEEE Network*, vol. 7, no. 5, pp. 8–18, September 1993.
- [29] S. Shenker and L. Breslau, “Two issues in reservation establishment,” in *Proc. of the ACM SIGCOMM '95*, October 1995, pp. 14–26.
- [30] C. Perkins, “IP mobility support,” Internet Request For Comments, October 1996, RFC 2002.

This page intentionally left blank

Chapter 11

PORATABILITY ARCHITECTURE FOR NOMADIC WIRELESS INTERNET USERS AND SECURITY PERFORMANCE EVALUATION

Mustafa M. Matalgah¹, Jihad Qaddour², Omar S. Elkeelany³ and Khurram P. Sheikh⁴

¹ University of Mississippi : ²Illinois State University : ³University of Missouri, Kansas City : ⁴Sprint

Abstract A value added service to broadband wireless network is the remote access virtual private network (VPN). The corporate legitimate portable users can connect to their offices through a wireless network from different locations and get secure services as if they were connected to the corporate local area network. One of the main challenges is to block illegitimate wireless users' requests. Registration and authentication functions should be implemented with highly secured wireless connection. These functions are accomplished by tunnelling the user information in a secured form to the corporate authentication server through the Internet traffic. The Corporate Authentication Server then grants or denies the user access. This chapter addresses various portability scenarios, architectures, implementation, and requirement issues for portable wireless Internet access systems. Moreover, performance evaluation and comparison are presented for the state-of-the-art security and authentication techniques.

Keywords: Wireless Internet, portability architecture, remote wireless authentication and security, VPN tunnelling.

1. Introduction

One of the most important requirements for high-speed fixed wireless Internet services is to support customer premises equipment (CPE) self install feature. This provides the user the ability to move around his location and re-establish radio and Internet protocol (IP) sessions easily. This is the simplest form of portability that can be supported to nomadic

users. In mobility it is required to keep the radio and IP sessions connected continuously. Whereas, for portability when a user moves from one regional base station to another then a hard handoff is experienced between the two regional base stations that belong to two different radio network controllers (RNC) in European networks (or mobile switching centers in American networks (MSC)). Different handoff scenarios and requirements need to be defined for both IP session handoff via a routing gateway, and radio network (RN) handoff via the radio link controller (RLC). If the user moves from one base station to another that belongs to the same RNC or MSC, then cell reselection need to be accomplished to reconnect the portable device at the new location. In mobility, cell reconnection for this scenario is performed smoothly without reestablishing a new session with the new cell. However, a new IP address reassignment might need to be established if subnet has changed.

When a portable device moves from one base station to another that belongs to a different RNC or MSC, then a new radio link protocol (RLP) signalling session with the new base station is needed for cell reselection. User registration and authentication is also needed if each RNC has its own server for Remote Authentication Dial In User Service (RADIUS). This server is used for dial-up security management.

This chapter addresses the following : 1) End-to-end element functions that meet portability requirements. 2) Various portability scenarios, architectures and implementation issues in fixed wireless Internet access systems. One of the major issues to be addressed is how to complete a secured time-bounded authentication, authorization and accounting when a user travels from one wireless market area to another through an IP network. 3) Private end-to-end architecture between the home and visited market network to complete a secured time-bounded portability requirement.

2. System Authentication

When a portable device moves from one base station to another that belongs to a different RNC, two methods are proposed for user registration and authentication (assuming each RNC belongs to a different RADIUS). In one method, since a different RADIUS serves each RNC, the new RNC will verify the user eligibility through its local (primary) RADIUS database, using tunnelling and routing protocols, for registration and authentication. The local RADIUS will then send interrogation signalling messages to all the RADIUSs in the network to verify the eligibility of that user. The RADIUS with the database that the user belongs to, will then reply with the eligibility status of that user to ac-

cept or deny access. Interrogation messages can be established through layer 2 tunnelling protocols such as layer 2 tunnelling protocol (L2TP) if private leased lines are utilized or layer 3 IP tunnelling protocols such as IPSec if public IP network is encountered. Tunnelling protocols will be covered in more details later in this chapter.

In the second method, it is assumed that there is secondary RADIUS in the network of the service provider that contains a copy of the entire primary RADIUSs at all the RNCs networkwide. This RADIUS might be co-located in one of the RNCs and connected to all the other RNCs in the network through secure private lines. Unlike the first method, the local (primary) RADIUS in the visited region will need to send authentication messages only to the central (secondary) RADIUS for user registration and authentication. This process saves the burden of interrogation messaging forwarding to all the RNCs networkwide.

In both methods, after the customer eligibility is confirmed, the new RNC will grant the user the authority to establish RLP and IP sessions through the access base station. The user information is then saved in the visited RADIUS database cache memory for future authentication. All the signaling messages and transmissions between and within the RNCs and RADIUSs may be established through the service provider IP network. A tunnelling protocol with security protocol such as a combination of L2TP and IPSec transport mode or IPSec tunnel mode might then be needed to ensure security. If leased private lines are used for transport, then L2TP alone is enough for delivering the messages through a point-to-point (PPP) [25] session.

3. Portability Architecture Elements

A high level architecture for a wireless Internet access network with the necessary elements to support portability is depicted in Figure 11.1. The definition and functionality of each of the network elements in this architecture are described in the following sections.

3.1 CPE

Customer premises equipment (CPE) is required to monitor few neighbor lists and measure the strongest power among them. The CPE then can send its measurements messages to the current serving base transceiver station (BTS), which will send a request message to the radio access controller for a hard hand-off. This request message is sent through a BTSs Broadcast Channel (BCCH). The user will then experience a disconnection and will be asked to reconnect. No authentication is required in this case if the two BTSs belong to the same RADIUS

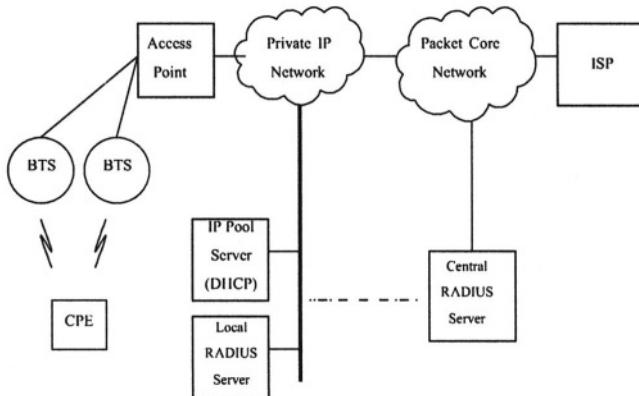


Figure 11.1. Network elements for portable wireless Internet architecture

server. A routing card at the radio controller is required to route the user information to the new sector or BTS through an L2TP session.

3.2 BTS and Radio Access Controller

Base transceiver station (BTS) should have the intelligence to sense user movements in the RF coverage area and between BTSs and inter-operate with the CPE and the radio controller portability requirements. Portable wireless Internet architecture is expected to include a radio controller that is capable of performing bandwidth allocation/management and traffic scheduling. The radio controller might be a part of each BTS or a one common radio controller might serve several BTSs. Its performance is determined by the system capability of supporting QoS and maintaining same data rate per user as power is changing when the user changes his location. Adaptive modulation and coding technique is also very important in bandwidth allocation and traffic scheduling. The radio access controller need to interface with a packet radio services supporting node that will provide the packet session management and handoff between radio access controllers and associated (BTSs) for portable devices.

The following are major functions that should be implemented in radio controller system and the BTS for a portable architecture:

1. Detection and authentication of new subscribers.
2. Dynamic allocation of bandwidth to subscribers and routing.
3. Cell selection and reselection on the wireless access network.
4. User Prioritization.
5. User Bandwidth Distribution.
6. Overload Control.
7. User traffic and quality of service (QoS) scheduling based on users service level agreement (SLA).
8. Advanced medium access controller (MAC) to support real-time QoS-based applications.
9. Ability to provide DiffServ.
10. Minimized Latency.
11. Operation, administration, and management (OAM) configuration capability.

Other functions provided by the resource allocation mechanisms include:

1. Differentiate between users on different tiers of service.
2. Provide fair queuing within these tiers of service so that a particular user cannot dominate resources.
3. Allocate resource so that power control can work efficiently.
4. Accommodate users on the edge of the cell who will probably be on lower transport formats with lower rates and aim to provide them with a reasonable throughput whilst not excessively reducing the overall cell throughput.

Resource allocation parameters that could be operator controlled are:

1. Tier weighting parameter: This provides a means of discriminating between groups of users or tiers and providing different levels of service to each tier.

2. Tier target rate: This allows the operator to define an upper limit on the throughput that a tier can experience. This is so that if a single lower tier user is present he will only experience maximum throughput for his tier of service. This can be disabled during off-peak periods for example.
3. Fairness parameter: This can be set between two extremes, one which divides the resources equally at the expense of the throughput to users on lower transport formats and the other which divides the throughput equally at the expense of overall cell throughput. This parameter provides a means of providing service anywhere between these two extremes.
4. Maximum number of downlink codes: This provides a means of trading coverage with overall system throughput.
5. Allocation window length: This determines how far in advance an allocation can be made.

3.3 IP Pool Server (DHCP)

The main function of the IP Pool or Dynamic Host Configuration Protocol (DHCP) Server is to perform IP address reassignment as needed when, the user is moving regionally and nationally or for driving vehicles with portable devices.

3.4 Encryption or Tunnelling Client

An encryption or tunnelling client is required at the RNC or master hub to guarantee and secure the messaging packets transmission through the IP network to the secondary RADIUS for radio access device authorization when the user is moving regionally or nationally. Authorization, control, and registration are required to be established in a secured form. They are also required for the Internet service provider (ISP) authorization. Tunnelling can also be used for IP session hard handoff in ISP network when IP address reassignment is required for driving vehicles portable devices. For example, the maximum speed allowed to achieve portability IP session handoff in Universal Mobile Telecommunication systems (UMTS) is less than 30 km/hr.

Tunnelling options include IPSec, L2TP, ... etc. These protocols will be explained in more details later in this chapter. IPSec is an encryption protocol that encrypts all received packets to be transmitted to the remote location for authentication. No pre setup time is required for this protocol. The only time delay to be experienced is from the encryption

process for each packet. IPSec is not a tunnelling protocol, but provides secured encrypted IP packet transmission when used in association with a tunnelling protocol such as L2TP. Further details about IPSec different modes are provided later in this chapter. L2TP is a tunnelling protocol to transmit PPP frames at Layer 2. A pre setup time is required to create the tunnel. TCP control channel (Layer 4) is used in the initial set up for the tunnel creation. After the tunnel creation is complete the PPP frames are transmitted over the IP cloud in which the nodes function as an intermediate hop switching (Layer 2). Tunnelling in L2TP is equivalent to Virtual Path Identifier (VPI) in ATM switching. L2TP can be used as an extension to the PPP connection from the CPE all the way to the ISP router or the RADIUS. More details are provided in the following sections of this chapter.

3.5 Re-encryption or Tunnelling Server

A re-encryption or tunnelling server is required at the remote RADIUS location to complete the tunnelling process or re-encrypt the messaging packets at the remote authentication server (RADIUS) for the radio network authorization or the remote ISP server for ISP authorization.

3.6 IP Network

If the authentication/authorization messaging encrypted packets or tunneled frames have to be transmitted through an Intranet or Internet, then it is expected to support the encryption or tunnelling requested by the client.

3.7 Local RADIUS Server

A local (or primary) RADIUS [32] server is required in each service market and keeps the database for all customers that belong to a regional service market. Each RNC (or MSC) is expected to belong to a RADIUS server for the registration and authentication.

3.8 Central RADIUS Server

It is expected to have a one central (or secondary) RADIUS server in the operators network that will be used to authenticate portable users (as explained in the second method in Section 2). The central RADIUS server provides backup for the local RADIUS and a master database for authentication of roaming portable users. When a user moves outside his home location, the radio access controller at the visited region will handle

the authentication and authorization through the secondary (central) RADIUS server.

4. Key Factors in Portability Architecture

The following are key factors that should be considered in designing portability network architecture:

4.1 QoS Management

The system should be able to adapt and identify users who are portable in nature. This is necessary on account of the fact that portable users are treated differently than fixed users. The primary reason for this is that data rates vary from one point to another in a given cell coverage area. This is driven by the ratio of the carrier to interference (C/I) distribution. Hence, it is critical that a given users' SLA changes with the available data rates in that region of the cell. Also, optimal market penetration should be maintained in the region of lowest data rates. Hence, the MAC layer should be capable of identifying such users and updating their traffic profile accordingly.

4.2 Service Level Agreement

Since a portable user always changes his/her location, it is not possible to guarantee high data rates for that user. This is primarily because of the fact that any given system which supports adaptive modulation supports data rate based on the C/I ratio at a given location in the cell. Hence, the SLA for a portable user should be less than that of a fixed wireless case, such that a large user population can be supported at the lowest modulation scheme.

4.3 Indoor Penetration

The air interface technology used by the portable network should be optimized for high downloading speeds to the user for carrying Internet Protocol (IP) traffic for Internet access and other portable/ fixed data applications operating in a non line-of sight environment with building penetration. Indoor wall penetration loss in the range 10 to 18 dB should be considered, depending on the technology, for cell design. In a typical network designed with sufficient link margin for building-penetration, the CPE Modem should operate indoors without the requirement for an external antenna.

4.4 Carrier to Interference (C/I) Ratio

The technology transport formats with less redundancy (weaker FEC, higher user data rate) require higher C/I than those with more redundancy (stronger FEC, lower user data rate) in order to meet a specified optimal value for the BLER. The basic rule is that if the block error rate (BLER) meets the specified optimal value, then maximum throughput is obtained. On the other hand, throughput is reduced if either BLER is more than the optimal value and then there are too many retransmissions to occur or if BLER is less than the optimal value and then FEC coding is more than necessary. Simulation model for the cell area coverage as related to C/I in the downlink should be provided in order to achieve an optimal cell design in non-line-of-site (NLOS) environments. Transport formats are defined by the technology in terms of different parameters such as number of codes, modulation type, turbo coding rate, puncturing limit, number of slots, ... etc. For each transport format using the ratio of the bit energy to noise density (E_b/N_0) requirement and the Processing Gain (G_p), the required C/I to achieve certain data rate can be calculated and then the coverage percentage is determined from the simulation model relating the C/I to the percentage cell coverage area. Figure 11.2 shows an example of a C/I cumulative distribution function (CDF) in a code division multiple access (CDMA) air interface system using a standard 19-cell cluster model with user population density of 3395 user/km^2 , 0.75 km radius, 100% frequency reuse, and a chip rate of 3.84 Mcps (in a 2.6 GHz spectrum with 5 MHz channels) [17].

The simulation model and path loss parameters used for calculating the C/I CDF in Figure 11.1 are summarized in Table 11.1

Table 11-1 Simulation parameters

| | |
|-------------------------------|----------------|
| Carrier frequency | 2600 MHZ |
| Chip Rate | 3.84Mcps |
| Frequency Reuse factor | N=1 |
| Pathloss Model | 3GPP TS25.942 |
| In-Building Loss | 15 dB |
| Cell Radius | 750 m |
| Subscriber Density | 3395 per sqkm |
| Log-Normal Standard Deviation | 10 dB |
| Antenna Front-to Back Ratio | 30 dB |
| Sectorization | Tri-Sectorcd |
| Antenna Pattern | $\cos(\theta)$ |
| Node-B Max Transmit Power | 34 dBm |
| Node-B Antenna Gain | 18 dBi |
| Antenna Height | 25 m |
| Node-B Tx Feeder Loss | 2 dB |
| Modem Antenna Gain | 2 dBi |
| Modem Receiver Noise Figure | 5 dB |

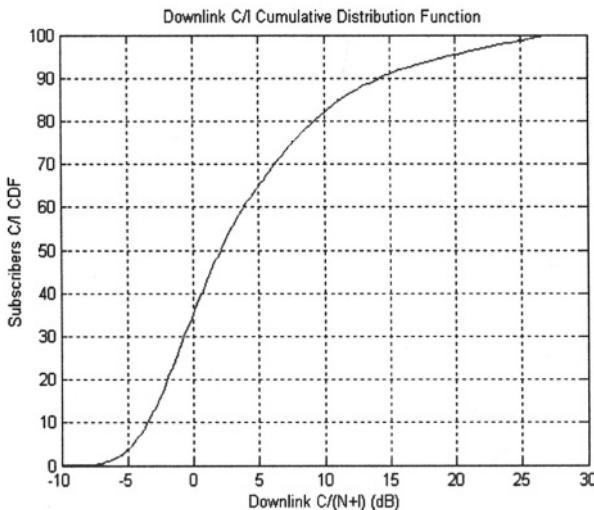


Figure 11.2. Downlink C/I CDF for a population density of 3395 users per square kilometer

5. Portability Scenarios

5.1 In-House Portability

For the in-house portability, the user can move around the house without shutting down the IP session. The house and the surrounding area are most likely assumed to be under same base station coverage. In this scenario, the following factors are considered:

1. CPE is initially provisioned with the best in-house QoS.
2. While the user is moving around the house a soft QoS will be required to keep certain data rate per the customer SLA. This will require maintaining a high SLA for the customer and providing the adaptive coverage per the SLA as the user moves. Adaptive coding and adaptive modulation techniques such as QPSK, 8PSK, 16QAM, or 64QAM are used for providing coverage adaptation.

Figure 11.3 shows a scenario for in-house portability sector architecture where the user is moving from point P1 to point P2 within sector 2.

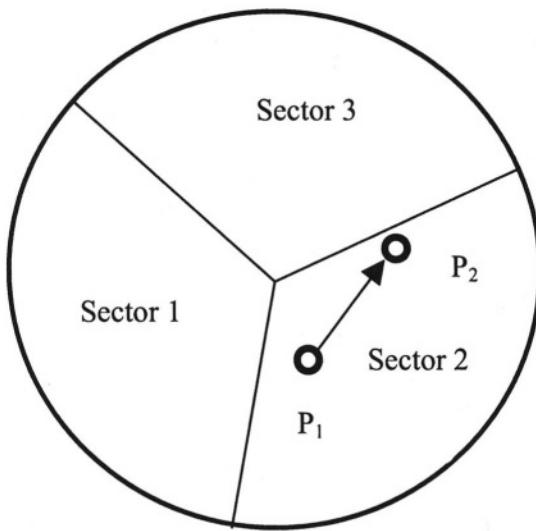


Figure 11.3. Intra-Sector probability architecture

5.2 Inter-Cell Regional Portability (Intra-RNC)

A higher level of regional portability is defined as when a user travels from one BTS to another BTS that belong to the same RNC as shown in Figure 11.4. This is referred to as Inter-Cell or Intra-RNC portability. If the user travels from one BTS to another BTS that belongs to the same RNC, then cell reselection should be established through the BTS and the RNC to establish a new RLP session with the new cell. In this case a new IP address needs to be reassigned (if subnet has changed). In this form of portability the following factors should be considered carefully.

In addition to the QoS, SLA, C/I and indoor penetration requirements mentioned above, virtual IP tunnelling and maybe VPN virtual path network)application will play a major role in this scenario. More details on tunnelling and security applications for portability are provided later in this chapter. Figure 11.4 shows a block diagram for inter-cell portability scenario architecture. The portable user is moving from point (P1) to point (P2) as shown in the figure.

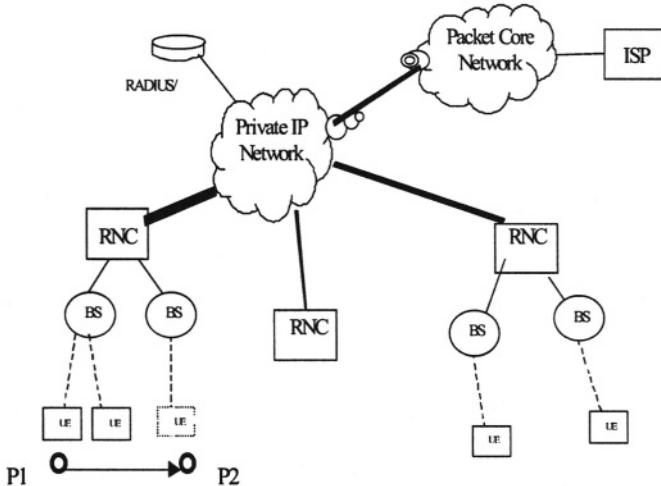


Figure 11.4. Inter-cell architecture for cell-to-cell regional portability within an RNC.

5.3 Inter-RNC Regional Portability (Same RADIUS)

Inter-RNC regional portability occurs when a user travels between two base stations that belong to different RNCs with same RADIUS. In this scenario, cell reselection is accomplished through the new BTS and the two RNCs to complete the RLP signaling process and establish a radio network (RN) session to initiate an IP session. A new IP address is assigned through the DHCP in the new RNC.

This is an on-network value-added hard handoff feature that needs authentication across the network between different RNCs. Figure 11.5 shows an architecture scenario for Inter-RNC regional portability. The portable user is moving from point (P1) to point (P2) as shown in the figure.

5.4 National Portability (Inter-RADIUS, On-Network, same technology)

National portability is the most complicated scenario. It can also be referred to as Inter-RADIUS portability. When portable device moves

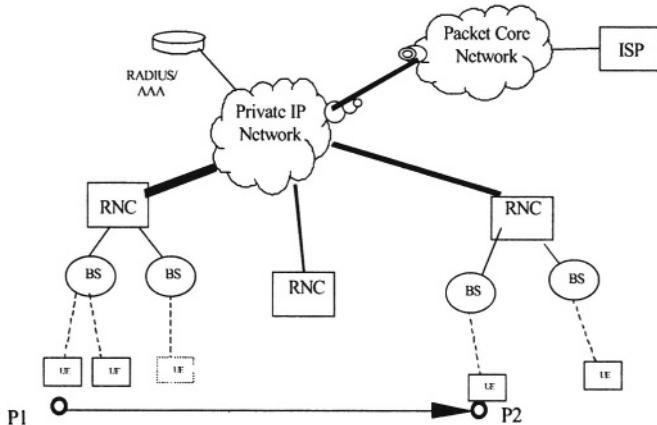


Figure 11.5. Inter-RNC regional portability architecture.

from one cell to another cell in a different RNC that belongs to a different RADIUS then cell reselection signalling and authentication/authorization process will need to occur across the network between the two RADIUSs. In this case the portable device needs to establish a new RLP session with the new cell via the BTS and the RNC. In this scenario the new RNC will check the user eligibility through its local RADIUS database, through tunnelling and routing protocols, for registration and authentication. Then using method I (refer to section 2), the RADIUS sends interrogation signalling messages to all RADIUSs across the network to check eligibility authentication of that user. The home RADIUS that the user belongs to, replies with the eligibility status of that user to confirm authentication and registration. Another authentication scenario can also be used (refer to method II in section 2).

Interrogation messages can be established through layer 2 tunnelling protocols such as L2TP, PPTP, and L2F or layer 3 IP tunnelling protocols such as IP in IP, GRE and IPSec. Other tunnelling mechanisms can be established using MPLS VPNs.

After the customer eligibility is confirmed, the new RNC provides the user the authority to establish an IP session through the visited access base station.

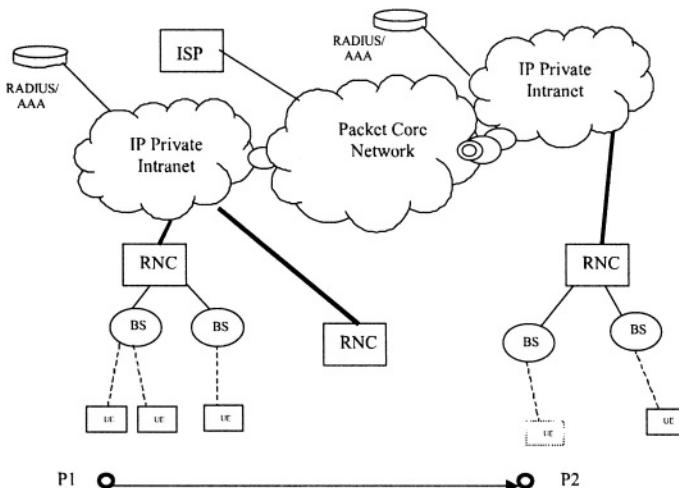


Figure 11.6. Inter-RNC architecture for national portability.

All the signaling messages and transmissions between the RNCs and the RADIUSs may be established through a private IP network if available between the two RADIUSs. A security protocol such as IPSec can be used to ensure security if public IP network is used. National portability architecture is shown in Figure 11.6.

5.5 Ubiquitous National Portability (OFF-Network, alternative carrier operator)

This occurs between two different service provider operators or within a service provider operator but with dual band technologies such as PCS (Personal Communications Service in the 1.9 GHz spectrum) and MMDS (Multichannel Multipoint Distribution Service in the 2.6 GHz spectrum). There are three types of roaming for ubiquitous portability:

1. Roaming from one MMDS service provider network to another MMDS service provider network. This is referred to as single-band dual-mode devices. Network architecture is illustrated in Figure 11.7 (Roaming from P1 to P2).
2. Roaming within the same service provider network but between two different services such as MMDS and PCS. This is referred

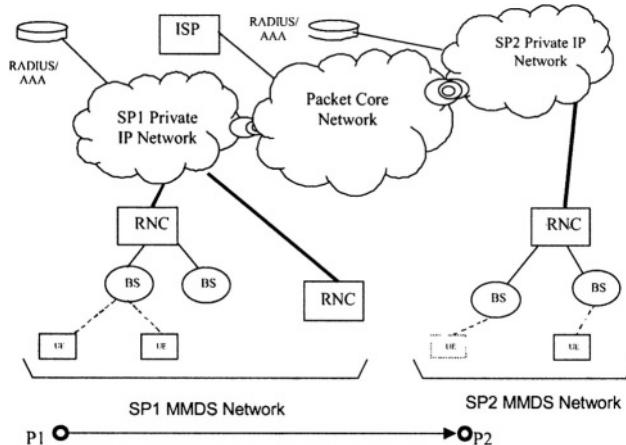


Figure 11.7. Architecture for single-band dual-mode ubiquitous national portability (SP1: Service Provider 1, SP2: Service Provider 2)

to as dual-band dual-mode devices. Network architecture is illustrated in Figure 11.8 (Roaming from P1 to P2).

3. Other roaming scenarios might also be practical such as dual-band dual-mode roaming between two different service providers such as roaming from MMDS services from one operator to PCS services from another operator or vice versa.

User interaction and/or pre-subscription are required for this scenario. By FCC (Federal Communications Commission) regulations the user must be informed of this form of roaming and the new charges rate.

6. IP Tunnelling

Tunnelling is a way of establishing a logical (virtual) path in an inter-networking infrastructure to transfer data from source to a destination on the network per an SLA. In a tunnelling protocol, data packets or frames are encapsulated in a new header that contain routing information such as source, destination, QoS level, data rate, ... etc. The virtual (logical) path through which the encapsulated packets travel through the network is called a “tunnel”

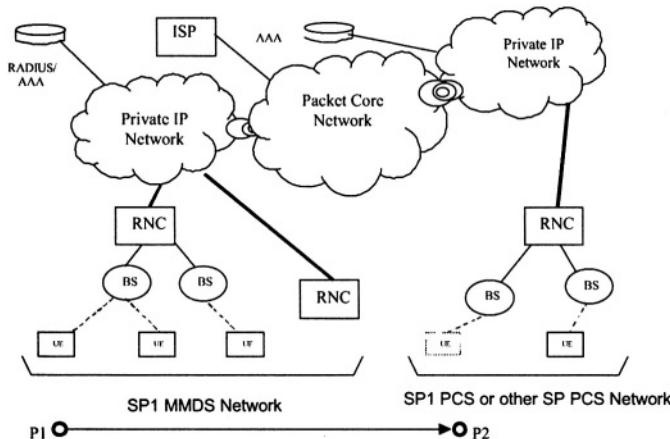


Figure 11.8. Architecture for dual-band dual-mode ubiquitous national portability (SP1: Service Provider 1, SP: Service Provider)

6.1 Tunnelling vs. Source Routing

Internet protocol (IP) source routing is a best effort and non-QoS transport protocol. IP source routing involves defining the path that the IP packet must follow through the network. Unresolved issues related to security, packet delay, packet drop, jitter and other performance metrics are still exist in IP source routing. These issues include examples such as non-guaranteed performance delivery by the IP router in forwarding datagrams for the IP source route options, incorrect processing of IP source route options, exclusion of such packets by firewalls, ... etc.

In IP tunnelling a logical path is set up to guarantee ... etc. In IP tunnelling a logical path is set up to guarantee packets delivery. QoS mechanism is to be utilized along with tunnelling to guarantee certain delivery performance requirements such as data rate (or throughput), time delay, jitter, ... etc.

6.2 IP Tunnelling for Portability Architecture

If a portable node moves from one RNC to another that belongs to different RADIUS, then authentication and registration need to be processed at the new RNC with a specified time interval. Tunnelling along

other QoS and Security mechanisms provide a secure and time bounded mechanism for this application. If the tunnel between the two RADIUSs has to go through a public IP/ATM network, then a SLA has to exist between the wireless Internet service provider and the IP/ATM service provider.

The main reason for utilizing tunnelling techniques in portability applications is to achieve the customer authentication and registration in a time bounded delay.

6.3 Layer 2 Tunnelling

Layer 2 tunnelling protocols are based on the well-defined point-to-point protocol (PPP) over L2TP and hence they inherit a suite of useful features like user authentication, dynamic assignment of addresses, data compression and encryption, and management. Layer 2 protocols correspond to the data link layer of the OSI seven layers model and they use frames as their unit of exchange. PPTP, L2TP, Layer 2 Forwarding (L2F) are examples of layer 2 protocols and they encapsulate the payload in a PPP frame to be sent across an internetwork.

6.3.1 L2TP (RFC 2661) . Remote system or L2TP Access Concentrator (LAC) client initiates a PPP connection across the PSTN cloud to the LAC. The LAC then tunnels the PPP connection across the Internet, Frame Relay, or ATM cloud to LNS whereby access to a Home LAN is obtained. Figure 11.9 shows a general schematic diagram for the L2TP topology [28]. L2TP utilizes two types of messages: control messages and data messages. Control messages are used in the establishment, maintaining and clearance of tunnels and calls. Data messages are used to encapsulate PPP frames being carried over the tunnel. L2TP is therefore considered a signalling protocol as well.

In L2TP protocol, control messages utilize reliable control channel (acknowledgement and retransmitting lost packets is supported) whereas the data channel is unreliable (acknowledgement and retransmitting lost packets is not supported). The following are the main features of L2TP:

1. Layer Two Tunnelling Protocol (L2TP) is a combination of PPTP and Layer 2 Forwarding (L2F), a technology proposed by Cisco Systems, Inc.
2. L2TP combines the best features of both PPTP and L2F.
3. L2TP encapsulates PPP frames to be sent over IP, X.25, Frame Relay, or ATM networks.

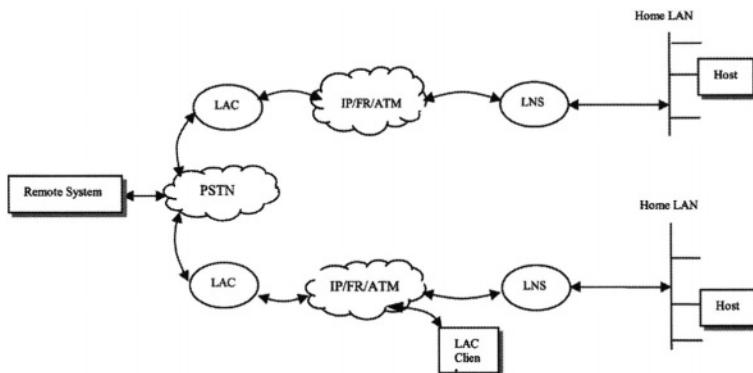


Figure 11.9. L2TP Generic Topology

4. When sent over an IP internetwork, L2TP frames are encapsulated as User Datagram Protocol (UDP) messages.
5. L2TP is used for tunnelling over the Internet as well as over private Intranets.
6. L2TP can be used for the establishment of compulsory as well as voluntary tunnels. Compulsory tunnels are tunnels setup by the LAC without the client knowing about it while voluntary tunnel setup is done by the client without the aid of the LAC as a separate entity. The client has the L2TP software running on it natively.

There are two tunnelling types associated with L2TP tunnelling protocols, Voluntary and Compulsory Tunnelling.

In voluntary tunnelling, a client directly connected to the Internet and running L2TP natively can also participate in tunnelling to the corporate LAN without the use of a separate LAC. A virtual PPP connection is created and the local L2TP client software creates a tunnel to the LNS. The client should have the tunnelling software installed and it functions as the end point of the tunnel. A block diagram that illustrates voluntary tunnelling is shown in Figure 11.10.

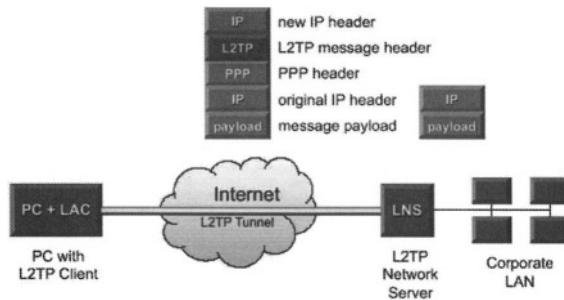


Figure 11.10. Voluntary tunnelling architecture.

In compulsory tunnelling, the remote user (PC) initiates a PPP connection across the PSTN cloud to the LAC. The LAC then tunnels the PPP connection across the Internet, Frame relay or ATM cloud to the L2TP Network Server (LNS) whereby access to the corporate LAN is obtained. The PC gets addresses from the corporate LAN via PPP-NCP negotiation. AAA services are provided by LANs management domain as if the user were connected to the LNS directly. This is illustrated in Figure 11.11.

6.3.2 PPTP (RFC 2637) . Point-to-Point Tunnelling Protocol (PPTP) defines a protocol, which allows the PPP to be tunnelled through an IP network [14]. There are two parallel components of PPTP. One component is a control connection between each PAC-PNS pair operating over TCP. The second component is an IP tunnel operating between the same PAC-PNS pair, which is used to transport generic routing encapsulation (GRE), encapsulated PPP packets for user sessions between a pair. The most commonly used protocol for remote access to the Internet is point-to-point protocol (PPP). PPTP builds on the functionality of PPP to provide remote access that can be tunnelled through the Internet to destination site. As currently implemented, PPTP en-

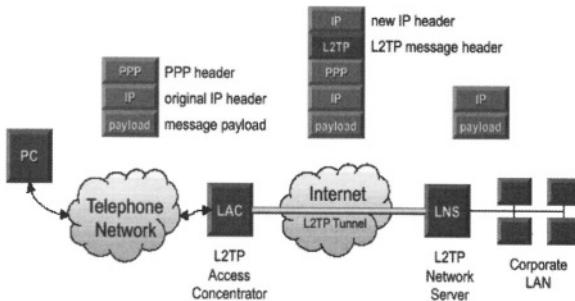


Figure 11.11. Compulsory tunnelling architecture.

encapsulates PPP packets using a modified version of GRE protocol, which gives PPTP the flexibility of handling protocols other than IP, such as Internet packet exchange (IPX) and network basic input/output system extended user interface (NetBEUI).

PPTP is one of the first protocols deployed for VPN. It has been a widely deployed as a solution for dial-in VPNs since Microsoft included support for it in RRAS for Windows NT Server 4.0 and offered a PPTP client in a service pack for Windows 95. Microsoft's inclusion of a PPTP client in Windows 98 practically ensures its continued use for the next few years, although it is not likely that PPTP will become a formal standard endorsed by any of the standard bodies (like the IETF). Because of its dependence on PPP, PPTP relies on the authentication mechanisms within PPP namely password authentication protocol (PAP) and CHAP. Because there is a strong tie between PPTP and Windows NT, an enhanced version of CHAP, MS-CHAP, is also used, which utilizes information within NT domains for security. Similarly, PPTP can use PPP to encrypt data, but Microsoft has also incorporated a stronger encryption called Microsoft point-to-point encryption (MPPE) for use with PPTP.

Aside from the relative simplicity of client support for PPTP, one of the protocol's main advantages is that PPTP is designed to run at open systems interconnection (OSI) Layer 2 as opposed to IPSec, which runs at Layer 3. By supporting data communications at Layer 2, PPTP can transmit protocols other than IP over its tunnels. PPTP does have some limitations. For example, it does not provide strong encryption for protecting data nor does it support any token-based methods for authenticating users.

6.3.3 L2F (RFC 2341). Layer Two Forwarding (L2F) Protocol, a technology proposed by Cisco [31], is a transmission protocol that allows dial-up access servers to frame dial-up traffic in PPP and transmit it over WAN links to an L2F server (a router). The L2F server then unwraps the packets and injects them into the network. Unlike PPTP and L2TP, L2F has no defined client. Note that L2F functions in compulsory tunnels only.

6.3.4 GRE (RFC 1701). A number of different proposals currently exist for the encapsulation of one protocol over another protocol. Other types of encapsulations have been proposed for transporting IP over IP for policy purposes. Generic Routing Encapsulation (GRE) is a general purpose encapsulation protocol and it is the attempt of this protocol to provide a simple, general purpose mechanism which reduces the problem of encapsulation from its current $O(n^2)$ size to a more manageable size [5]. In the most general case, a system has a packet that needs to be encapsulated and delivered to some destination. We will call this the payload packet. The payload is first encapsulated in a GRE packet. The resulting GRE packet can then be encapsulated in some other protocol and then forwarded. We will call this outer protocol the delivery protocol. Overall packet format is explained in Figure 11.12.

In forwarding GRE packet, normally a system, which is forwarding delivery layer packets, will not differentiate GRE packets from other packets in any way. However, a GRE packet may be received by a system. In this case, the system should use some delivery-specific means to determine that this is a GRE packet. Once this is determined, the Key, Sequence Number and Checksum fields if they contain valid information as indicated by the corresponding flags may be checked. If the Routing Present bit is set to 1, then the Address Family field should be checked to determine the semantics and use of the SRE Length, SRE Offset and Routing Information fields. The exact semantics for processing a SRE for each Address Family is defined in other documents. Once all SREs have been processed, then the source route is complete, the GRE header

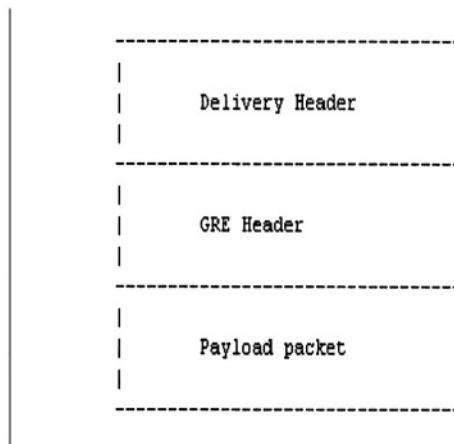


Figure 11.12. GRE packet format.

should be removed, the payload's TTL must be decremented (if one exists) and the payload packet should be forwarded as a normal packet. The exact forwarding method depends on the Protocol Type field.

6.4 Layer 3 Tunnelling

6.4.1 IPSec Tunnel (RFC 2401) . IPSec is a layer 3 solution for secured tunnelling, which means authenticated and encrypted tunnel [11]. It provides security (Authentication and encryption) at Layer 3 (IP Network Layer) by enabling a system to select required security protocols, and determine the algorithm(s) for encryption. IPSec has two security protocols, IP Authentication Header (IPSec AH) [9], and IP Encapsulating Security Payload (IPSec ESP) [10]. IPSec provides two types of security algorithms: symmetric encryption algorithms (e.g. Data Encryption Standard DES) [29], [30], [15], [27] and, one-way hash functions (e.g., Message Digest MD5, Secured Hash Algorithm SHA-1) [21], [14], [13], [24].

IPSec is used to secure tunnels against false data origins, and encrypt traffic against unwanted network passive or active intruders from listening or modifying actions. IPSec can be used between pairs of Security Gateways, pairs of peer hosts or, combined (e.g. between a host and

corporate equipment through security Gateways.) It has two modes of operation: Transport mode, and Tunnel mode. The transport mode provides upper layer protection (transport layer). Transport mode applies to pairs of peer hosts. Also in this mode, traffic goes to the ultimate destination directly. In the Tunnel mode, a tunneled IP protection is provided (via Gateways). Traffic have to pass through a gateway (i.e. firewall) to access ultimate destination.

In the following sections, we present terminology used in the IPSec context then present the two types of IPSec AH and ESP.

A. Terminology.

- Security Association (SA): a Simplex (unidirectional) secured connection identified by the following:
 1. Security Parameter Index (SPI), two bytes saved in IPSec header, which with the following uniquely identifies the SA.
 2. IP Destination Address from the IP packet header.
 3. Security protocol identifier (50 for AH and 51 for ESP.) One byte saved in the IP packet header.
- Integrity: there are three types of integrity in the context of IPSec:
 1. Connectionless integrity, that detects modifications of a single IP packet by the use of an algorithm specific Integrity Check Value (ICV).
 2. Anti-replay integrity, that detects duplicate IP packets at the receiver end by the use of sequence numbers in IPSec header.
 3. Connection-oriented integrity, that detects lost or reordered IP packets at the receiver end by the use of sequence numbers in IPSec header as well.

B. IPSec Authentication Header (AH). IPSec Authentication Header (AH) provides Connectionless Integrity, Data Origin Authentication, and Anti-Replay Integrity (optional - not enforced at receivers end). Figure 11.13 depicts the IP AH header format. The Next Header field is of 8 bits size and specifies the type of the transport protocol used in the upper layer. The Payload Length field is also an 8 bits size, and contains the IPSec header length in words (32 bit) minus 2 words (e.g. $3+3-2=4$, if Authentication Data is 3 words (96 bits)). The sender always transmits the Sequence Number field (32 bits), but receiver might

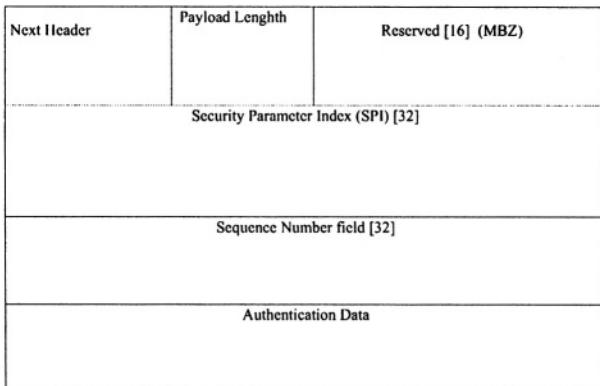


Figure 11.13. IPSec AH Header format.

optionally act on it. Finally, the Authentication Data field (variable size, multiple of 32 bits), contains Integrity Check Value (ICV) for the attached packet (Including the AH header itself). Reserved bits (Must Be Zero) are used for future extensions.

Figure 11.14 depicts the coverage of authentication protection for IP AH in Transport mode and Tunnel mode. Note how is the IP AH header is inserted between the IP Header (IPH1) and the upper header (e.g. TCP Header) in Transport mode. Note also, in tunnel mode the IPSec AH constructs another IP Header (IPH2) for the IP address of destination gateway.

The ICV is computed first at transmitter by the use of a common authentication algorithm that is also known to the receiver. Then ICV is recomputed at receiver and compared to match the received value for authentication integrity. ICV computation excludes non-predictable IP Header (IPH) fields like TTL, Flags, TOS, Fragment offset, Checksum, ... etc. If IP fragmentation occurs at sender, it should be performed after AH processing. The IP reassembly should then be performed before AH processing at the receiver.

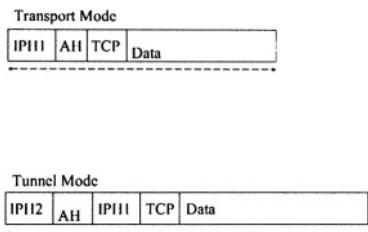


Figure 11.14. IP AH Authentication protection (dotted arrows)

C. IPSec Encapsulating Security Payload (ESP). IPSec encapsulation security payload (ESP) provides Confidentiality (Encryption), Connectionless Integrity (optional, not enforced at receiver end), Data Origin Authentication (optional, not enforced at receiver end), Limited Traffic Flow Confidentiality, and Anti-Replay Integrity. Figure 11.15 depicts the IP ESP header format. The Next Header field is of 8 bits size and specifies the type of the transport protocol used in the upper layer. The Pad length contains the number of pad bytes inserted by the encryption algorithm. The Payload Length field is also an 8 bits size, and contains the IPSec header length in words (32bit) minus 2 words (e.g. $3+3-2=4$, if Authentication Data is 3 words (96bits)). The sender always transmits the Sequence Number field (32 bits), but the receiver might optionally act on it. Finally, the Authentication Data field (variable size, multiple of 32 bits) contains Integrity Check Value (ICV) for the encapsulated packet and the ESP header/ trailer (not including the authentication data itself).

Figure 11.16 depicts the coverage of authentication and encryption protections for IP ESP in Transport mode and Tunnel mode. The ESP header is inserted exactly same way as in AH header. The ESP trailer is

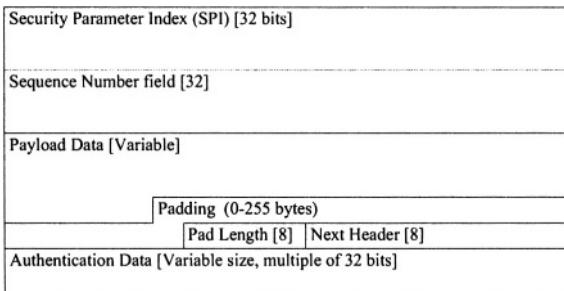


Figure 11.15. IPSec ESP Header format.

inserted after the payload data and before the ESP authentication data. ICV computation steps are the same as in IPSec AH.

6.5 MPLS Tunnelling

Multi protocol label switching (MPLS) is another technology that can be used in creating Virtual Private Networks (VPN) tunnels. Here, VPN tunnels are created using MPLS and IPSec in a public IP network. Virtually all the interfaces in an IP based network could be implemented via a MPLS VPN tunnels leased from a public carrier. The current efforts in IETF are focused on defining MPLS architecture and associated label distribution protocols. Label switching router (LSR) and label switching path (LSP) or tunnel are the main features of MPLS.

7. Performance Analysis of IPsec Protocol

IPSec provides two types of security algorithms, symmetric encryption algorithms (e.g. Data Encryption Standard DES) for encryption, and one-way hash functions (e.g., Message Digest MD5 and Secured Hash Algorithm SHA1) for authentication. This section presents performance analysis and comparisons between these algorithms in terms of time

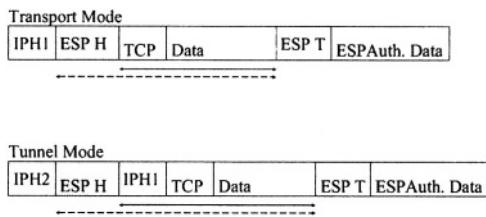


Figure 11.16. IP ESP Authentication protection (dotted arrow) and Encryption protection (solid arrow).

complexity and space complexity. Performance Parameters considered are processing power and input size. The analysis results revealed that HMAC-MD5 can be sufficient for the authentication purposes rather than using the more complicated HMAC-SHA1 algorithm. In encryption applications, authentication should be combined with DES. This section focuses on IPSec authentication and encryption space complexity, computation time complexity, and protocol limitations.

7.1 Space Complexity

IPSec AH/ESP header size is 12/10 bytes fixed header fields respectively, plus the variable size authentication data. Authentication data is algorithm and packet specific field. IPSec authentication uses Keyed-Hashing for Message Authentication (HMAC) [14], [13], [12] combined with Message Digest algorithm (MD5) [24], or Security Hash Algorithm (SHA-1) [13]. All algorithms use secret key distribution. In combined HMAC-MD5, the key size is 128 bit while in the combined HMAC-SHA-1 the key size is 160 bits. In both cases the algorithm works with 64-byte message blocks. They generate a truncated ICV of 96 bits (12 bytes) to conform with the IPSec AH and ESP authentication data size. This

means a total header size of 24/22 bytes for AH/ESP respectively, per packet. This overhead is needed for each authentication request transmitted through a security association in transport mode. Additional 20 bytes of IP header are also needed per packet for tunnel destination in tunnel mode. A total overhead of 44 bytes per packet assumes no header compression mechanisms used. Table 11.2 summarizes these results for both Transport and Tunnel modes.

IPSec ESP encryption may use extra padding bytes (0 to 255 bytes). They are algorithm and payload specific.

Table 11-2. IPSEC HEADER FIELD SIZES (BYTES)

| Protocol | IPSec AH | | | IPSec ESP | | |
|----------------|----------|-------------|----------|-----------|-------------|----------|
| Transport Mode | Fixed 12 | Variable 12 | Total 24 | Fixed 10 | Variable 12 | Total 22 |
| Tunnel Mode | 12+20 | 12 | 44 | 10+20 | 12 | 42 |

7.2 Time Complexity

7.2.1 Encryption. IPSec ESP Encryption algorithm is triple Data Encryption Standard (3DES) in Cipher Block Chaining (CBC) mode [15]. Encryption using DES algorithm is the most time consuming process. DES uses a 56-bit key, and block sizes of 64 bits. The algorithm has 19 distinct steps. The first step is a key independent transposition on the 64bit input block. The last step is the exact inverse of this transposition. In step 18, a 32 bit SWAP operation is performed. The remaining steps (2 to 17) are functionally identical and are dependent on different portions of the input key. Each of these 16 steps takes two 32 bit inputs, and produce two 32 bit outputs. The left output is a COPY of right input. The right output is an XOR of left input and a function f of right input and the step key. The function f consists of 4 operations. First, a 32:48 bit transposition/expansion of the right input is applied; second a 48 bit XOR of the output with the step key. Then a group mapping is performed to reduce the output size from 48 to 32 bits using two dimensional look-up table of 4 rows by 16 columns for all possible 64 inputs. Finally, 32-bit transposition is performed.

Each of the 16 steps has a special key, which is derived from the 56-bit key using the KS function. For the KS function, an initial 56-bit transposition is performed to the input key. Before each step, the key is divided into two 28- bit sub-keys. Each of which is rotated left using LEFT SHIFT operation. Finally, 56:48 bit transposition/reduction is performed.

Table 11.3 lists DES basic operations and their equivalent simple ones. It also gives space requirement for each operation. Table 11.4 lists the DES and 3DES computations of total number of simple operations needed. 3DES is the chained form of DES with a chain size of 3. For simplicity, 3DES is assumed to have only 3 times number of operations as DES has. 3DES requires an extra random Initialization Vector (IV) of 8 bytes. For complete details of DES and 3DES algorithms see [29] [30].

As shown in table 11.4, the total number of operations per 64-bit block is 8091 operations. Given a packet size of N bits then the number of blocks is given by

$$n = \lceil \frac{N}{64} \rceil \quad (11.1)$$

Where $\lceil \cdot \rceil$ means the smallest integer bigger than or equal to the operand. Therefore, 3DES time complexity is of $O(n)$.

Table 11-3. DES basic operations

| Basic Operation | Equivalent simple operation Type | #Times needed b | Space needed b |
|--|-------------------------------------|--------------------|-------------------|
| b bit transposition | One dimensional table look-up | | |
| Two dimensional table map (for 6:4 bit map.) | Multiply | 1 | |
| | Add | 1 | |
| | One dimensional table look-up | 1 | 4 rows×16 cols |

Table 11-4 DES Operation in one block encryption

| Step# | operation | #Times | Equiv.Total* | Notes |
|-------|---|--------|--------------|------------|
| 1,19 | 64 bit transposition | 2 | 64×2 | |
| 2-17 | 32 bit COPY | 16 | 16 | |
| 2-17 | 32 bit XOR | 16 | 16 | |
| 2-17 | 48 bit transposition | 16 | 48×16 | |
| 2-17 | 48 bit XOR | 16 | 16 | |
| 2-17 | 6:4 bit Two Dimensional Table Mapping | 8×16 | 3×128 | |
| 2-17 | 32 bit transposition | 16 | 32×16 | |
| 1 | 56 bit transposition | 1 | 56 | |
| 2-17 | 28 bit LEFT SHIFT | 2×16 | 32 | |
| 2-17 | 48 bit transposition | 16 | 48×16 | |
| 18 | 32 bit SWAP | 1 | 1 | |
| | | | 2697 | Pre-out |
| | | | 8091 | DES Total |
| | | | | 3DES Total |

*using equivalent simple substitution from table 11.3

Figure 11.17 shows the computed encryption time in microseconds as function of input packet size in blocks of 64 bits and processing power in Millions of Instructions Per Second (MIPS). For example, time complexity of 3DES means approximately 4 Mbps encryption rate with a 500 MIPS.

7.3 Authentication

As presented in the previous section, IPSec Authentication algorithms are HMAC-MD5, or HMAC-SHA-1. These two algorithms are explained in this section.

7.3.1 MD5. The first step in MD5 algorithm is padding the original message by appending 1 to 512 bits to it. The original message size is also appended in 64 bits. The total padded message with the message size field should have a size, which is multiple of 512 bits (64 bytes).

Table 11-5 MD5 Operations per block

| Round # | Base Function ♠ | #Steps | Operations /step | Total |
|----------|--|--------|------------------|-------|
| 1 | F(X,Y,Z)= X [^] Y ^v -X [^] Z | 16 | 12 | 192 |
| 2 | G(X,Y,Z)= X [^] Z ^v Y [^] -Z | 16 | 12 | 192 |
| 3 | H(X,Y,Z)= X ^v Y [^] Z | 16 | 10 | 160 |
| 4 | I(X,Y,Z)= Y ^v (X [^] V-Z) | 16 | 11 | 176 |
| Total(T) | | | | 720* |

♠ Symbols \wedge for AND, \vee for OR, \neg for NOT and \oplus for EXCLUSIVE OR

* Plus 24 operations per block for initiation and termination

The algorithm uses four 4-byte registers (A, B, C, and D) which keep intermediate and, ultimately, the final value of the message digest of 128 bits. For each block of 64 byte of the message the algorithm performs 4 rounds of 16 steps each. In each step, one of the functions in table 11.5 is computed in terms the registers (A, B, C, or D), the current input block, and a lookup value. By investigating MD5, one can find out that 10 to 12 operations per step are needed. Thus, in table 11.5, the total number of operations (T) is shown as 744 (720 + 24) operations per block. The time complexity is $O(n)$, where n is number of input blocks given by

$$n = \frac{N}{512} \quad (11.2)$$

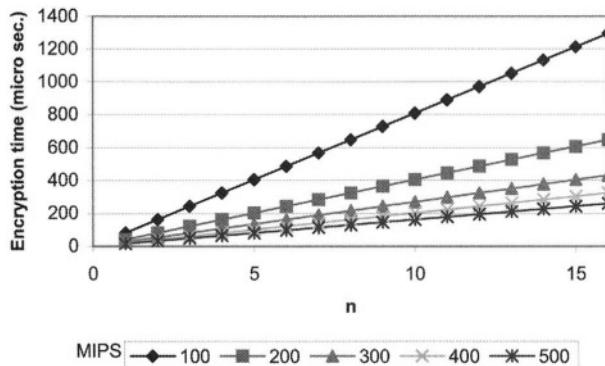


Figure 11.17. 3DES Encryption vs. no. of blocks and processing power.

$$N = X + pad + Size \quad (11.3)$$

where X is the input text, pad is the padding field, $size$ is the size field, and N is the total message size.

Figure 11.18 shows the computed algorithm computation time (digest time) in microseconds as a function of input size in blocks of 512 bits (64 bytes) and processing power in MIPS. The algorithm has been tested on a Pentium II machine with 300 MHz rate for one million 1024 byte messages. The digest time computed was 159 seconds. The digest rate then is 51.5 Mbps. This means that every operation counted in table 11.5 executed on the average in 6 cycles.

7.3.2 HMAC-MD5. The combined HMAC-MD5 algorithm is formulated as follows:

$$MD5(K_o, MD5(K_i, Text))$$

$$K_i = Key \oplus ipad \quad (11.4)$$

$$K_o = Key \oplus opad \quad (11.5)$$

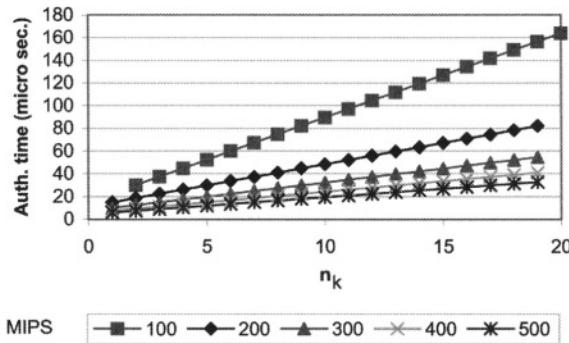


Figure 11.18. MD5 Digest time vs. no. of input blocks and processing power.

where K_i and K_o are two extended forms (512-bit) of the input Key and are generated by exclusive or the Key with $ipad$ the inner padding (512 bits), and $opad$ the outer padding (512 bits). Key is an arbitrary size secret key shared by sender and receiver. $Text$ is the given input message subject to authentication. For an input text of size X bits, the number of input blocks for the inner MD5, n_k , is

$$n_k = \frac{N}{512} \quad (11.6)$$

$$N = X + K_pad + size \quad (11.7)$$

where K is the size of the extra appended inner form of the key (512 bits).

In the outer MD5, the output of the inner MD5 (128-bit digest) is appended to K_o (512 bit). According to MD5, this is padded to two 512-bit blocks. Thus, the HMAC-MD5 total number of operations (T) is

$$T(n_k) = 32 + (2 + n_k) \times 744 \quad (11.8)$$

where 32 comes since in (11.4) and (11.5) the XOR operands are of size 512 bits. In a machine of word size 32 bits, this has to be partitioned in

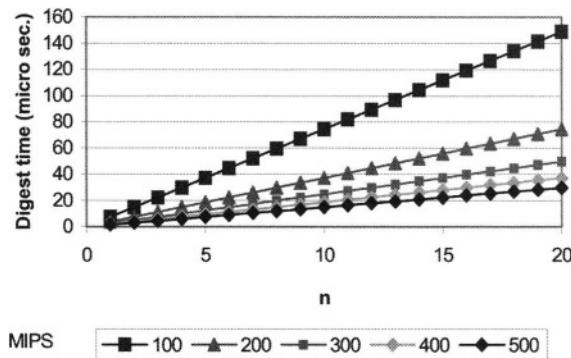


Figure 11.19. HMAC-MD5 Authentication time vs. no. of input blocks and processing power.

16 consecutive XOR operations. A total of 32 extra operations is added. The time complexity is then $O(n_k)$.

Figure 11.19 shows the computed authentication time in microseconds vs. no. of input blocks and processing power in MIPS.

7.3.3 SHA-1. SHA1 follows exactly the same way of message padding as in MD5 algorithm. However, the algorithm uses five 4-byte intermediate registers instead of four. Thus, the final value of the message digest is 160 bits. For each block of 64 bytes of the message the algorithm performs 4 rounds of 20 steps each. In each step, a functional computation based on the temporary registers, the current input block, and a constant value. By investigating SHA1, one can find out that 10 to 13 operations per step are needed. In table 11.6, the total number of operations (T) is computed as 1110 (=900+210) operations per block. The algorithms time complexity is $O(n_k)$, where n_k is the number of input blocks as in (11.6).

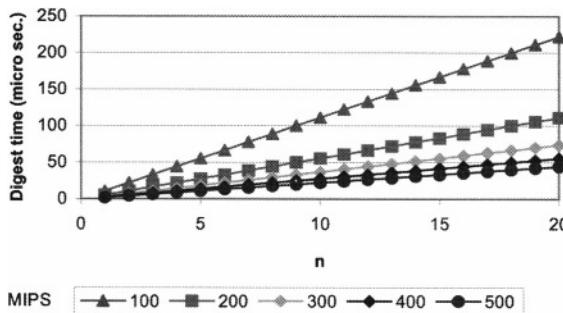


Figure 11.20. SHA1 digest time vs. no. of input blocks and processing power.

Table 11-6. SHA1 Operations per block

| Round# | Base Function | Steps | Operation/Step | Total |
|--------|--|-------|----------------|-------|
| 1 | $F(X,Y,Z)=X^A Y \vee \neg X^A Z$ | 20 | 12 | 240 |
| 2 | $G(X,Y,Z)=X \oplus Y \oplus Z$ | 20 | 10 | 200 |
| 3 | $H(X,Y,Z)=$ $X^A Y \vee X^A Z \vee Y^A Z$ | 20 | 13 | 260 |
| 4 | $I(X,Y,Z)=X \oplus Y \oplus Z$ | 20 | 10 | 200 |
| | | | Total | 900* |

* Plus 210 operation per block for initiation and termination

Figure 20 shows the computed digest time in microseconds as a function of number of input blocks and MIPS.

7.3.4 HMAC-SHA-1. The combined HMAC-SHA1 algorithm is formulated as follows:

$$SHA1(K_o, SHA1(K_i, Test))$$

The total number of operations (T) needed for HMAC-SHA1 is of $O(n_k)$ where,

$$T(n_k) = 32 + (2 + n_k) \times 1110 \quad (11.9)$$

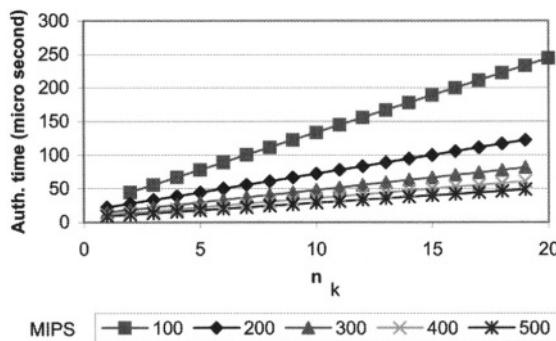


Figure 11.21. HMAC-SHA1 Authentication vs. no. of input blocks and processing power

Figure 11.21 shows the computed authentication time in microseconds as a function of number of input blocks and processing power in MIPS.

In summary Figure 11. 22 shows a comparison between 3DES, HMAC-MD5, and HMAC-SHA1 throughput vs. processing power. It shows how HMAC-MD5 supercedes HMAC-SHA1. It also shows how far the authentication throughput is as compared to encryption throughput.

7.4 Protocol Limitations

1. IPSec is IP based only. It does not provide security for any other network layer protocols.
2. Encryption using DES is based on a secret key of 56-bit size. In todays equipment this is very small key to break even if using brute force approach given the knowledge of plain text- cipher text pair. The addition of extra chaining of 3 levels (3DES) improves this limitation by the addition of an extra Initial Vector (IV) of random 64 bits. Practically, the IV can be set as the last 8 bytes of an encrypted packet for the next encryption process. In this case 3DES can be logically extended over consecutive packets [15]. Encryption should not be offered without the use of data origin

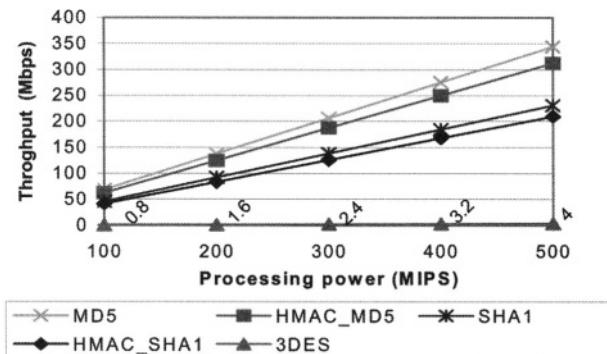


Figure 11.22. Throughput comparison vs. processing power.

authentication. After all, 3DES is much better than sending plain text.

8. Conclusions

In this chapter, we presented various portability scenarios, architectures, implementation, and requirement issues for nomadic wireless Internet users. Secured and fast connection for authentication and authorization is crucial in today's IP network. We focused on the state-of-the-art VPN tunnelling techniques that are required for remote nomadic wireless users authentication and authorization. IPSec provides two types of security mechanisms symmetric encryption algorithm (e.g., DES), and one way hash functions for authentication (e.g., MD5 and SHA 1). Performance evaluation and comparison between these mechanisms in terms of computation time complexity and space complexity have been presented in this chapter (parameters considered are power and input size). Our results revealed that HMAC-MD5 is sufficient for authentication purpose rather than using the more complicated HMAC-SHA1 algorithm. Moreover, MD5 is having higher throughput as compare to SHA1. Since the IPSec requires only 96 bits of the message digest, MD can be sufficient for the authentication purpose. 3DES en-

cryption throughput is very low compared to authentication throughput. It should only be used in critical user information not for regular traffic flow. Encryption if needed should be combined with authentication. In this case if the message fails authentication, decryption process is waived (not performed).

References

- [1] R. Droms, "Dynamic Host Configuration Protocol," IETF RFC 1541, October 1993.
- [2] O. S. Elkeelany, M. M. Matalgah, K. P. Sheikh, M. Thaker, G. Chaudhry, D. Medhi, and J. Qaddour, "Performance Analysis of IPsec Protocol: Encryption and Authentication," IEEE Communications Conference (ICC 2002), pp. 1164-1168, New York, NY, April 28 - May 2, 2002.
- [3] B. Fox and B. Gleeson, "Virtual Private Networks," IETF work, RFC 2685, September 1999.
- [4] K. Hamzeh, G. Pall, W. Verthein, J. Taarud, W. Little, G. Zorn, "Point-to-Point Tunneling Protocol (PPTP)," IETF RFC 2637, July 1999.
- [5] S. Hanks, T. Li, D. Farinacci, P. Traina, "Generic Routing Encapsulation (GRE)," IETF RFC 1701, October 1994.
- [6] Lin J. Heinanen, G. Armitage and A. Malis, "A framework for IP based Virtual Private Networks", IETF RFC 2764, February 2000.
- [7] Walter Honcharenko, Jan P. Kruys, David Y. Lee, and Nitin J Shah, "Broadband Wireless Access," IEEE communication magazine, Jan 1997.
- [8] International Business Machines Corporation, Using IPSEC to construct secure VPN, 1998. <http://www.firstvpn.com/papers/ibm/ipsecvpn.htm>
- [9] S. Kent and R. Atkinson, "IP Authentication Header," IETF RFC 2402, 1998.
- [10] S. Kent and R. Atkinson, "IP Encapsulating security Payload (ESP)," IETF RFC 2406, 1998.

- [11] S. Kent and R. Atkinson, “Security Architecture for the Internet Protocol,” IETF RFC 2401, 1998.
- [12] H. Krawczyk, M. Bellare, and R. Canetti, “HMAC: Keyed-Hashing for Message Authentication,” RFC 2104, February 1997.
- [13] C. Madson and R. Glenn, “The Use of HMAC-MD5-96 within ESP and AH,” RFC 403, November 1998.
- [14] C. Madson and R. Glenn, “The Use of HMAC-SHA-1-96 within ESP and AH,” RFC 404, November 1998.
- [15] C. Madson and N. Doraswamy, “The ESP DES-CBC Cipher Algorithm with Explicit IV,” RFC 2405, November 1998.
- [16] G. S. Malkin, “Dial-in Virtual Private Networks Using Layer 3 Tunneling,” IEEE Computer Society, 22nd Annual Conference on Local Computer Networks, 1997.
- [17] M. M. Matalgah, J. Qaddour, A. Sharma, and K. P. Sheikh, “Capacity Planning in UMTS WCDMA TDD: Throughput and Spectral Efficiency Analysis,” Proceedings of the 3GWireless’2003 Conference, San Francisco, CA, May 27 - 30, 2003.
- [18] G. McGregor, “PPP Internet Protocol Control Protocol,” IETF RFC 1332, Merit, May 1992.
- [19] G. Meyer, “The PPP Encryption Control Protocol (ECP),” IETF RFC 1968, June 1996.
- [20] P. Mockapetris, “Domain Names -Concepts and Facilities,” IETF RFC 1034, 1987 Conference, San Francisco, CA, May 27 - 30, 2003.
- [21] NIST, FIPS PUB 180-1: “Secure Hash Standard,” 1995.
- [22] B. Patel, B. Aboba, W. Dixon, G. Zorn, “Securing L2TP using IPsec,” Internet-Draft, July 2001.
- [23] D. Rand, “The PPP Compression Control Protocol,” IETF RFC 1962, June 1996.
- [24] R. Rivest, “The MD5 Message-Digest Algorithm,” RFC 1321, April 1992.
- [25] W. Simpson, “The Point-to-Point Protocol (PPP),” IETF RFC 1661, July 1994.
- [26] W. Simpson, “The Point-to-Point Protocol (PPP),” IETF RFC 1661, July 1994.

- [27] A. S. Tanenbaum, "Computer Networks," Prentice Hall PTR, 3rd Ed., pp. 588-595, 1996.
- [28] W. Townsley, A. Valenciam A. Rubens, G. Pall, G. Zorn, and B. Palter, "Layer Two Tunneling Protocol (L2TP)," IETF RFC 2661, 1999.
- [29] US National Bureau of Standards, "DES modes of operation," Federal Information Processing Standard (FIPS) publication 81, December 1980. <http://www.itl.nist.gov/fipspubs/fip81.htm>
- [30] US National Bureau of Standards, "Data Encryption Standard," Federal Information Processing Standard (FIPS) publication 46-2, December 1993. <http://www.itl.nist.gov/fipspubs/fip46-2.htm>
- [31] A. Valencia, M. Littlewood, T. Kolar, "Cisco Layer Two Forwarding (Protocol) L2F," IETF RFC 2341, May 1998.
- [32] Willens, S. A. Rubens, W. Simpson, C. Rigency, "Remote Authentication Dial In User Service," IETF RFC 2509, January 1997.
- [33] Willens, S. A. Rubens, W. Simpson, C. Rigency, "Remote Authentication Dial In User Service," IETF RFC 2138, April 1997.

This page intentionally left blank

Chapter 12

DESIGN AND IMPLEMENTATION OF A SOFTSWITCH FOR THIRD GENERATION MOBILE ALL-IP NETWORK

Vincent W.-S. Feng and Yi-Bing Lin

*Department of Computer Science & Information Engineering
National Chiao Tung University
Taiwan, R.O.C.*

S.-L. Chou

*Computer and Communication Research Laboratory
Industrial Technology Research Institute
Taiwan, R.O.C.*

Abstract This chapter describes the design and implementation of a softswitch for UMTS developed by Computer and Communication Research Laboratory (CCL) of Industrial Technology Research Institute (ITRI). This softswitch can be utilized as call agent (media gateway controller) for the third generation mobile all-IP network such as UMTS. The CCL Softswitch follows the reference architecture proposed by International Softswitch Consortium. In our approach, the Intelligent Network (IN) call model is implemented to interwork with existing IN devices. We design protocol adapter modules and service provider interface to ensure that multiple VoIP protocols can be supported without modifying the core of the softswitch. Furthermore, the message flows of call setup, call transfer and inter-softswitch call are described to show the feasibility of the softswitch.

Keywords H.323, Megaco, MGCP, SIP, softswitch, UMTS, VoIP

1. Introduction

Most of today's telecommunications networks are based on the circuit-switched technologies. Next generation networks such as the third generation (3G) mobile telecommunications systems [1, 2] utilize the packet-

switched technologies. Many new applications and services have been developed based on such technologies and networks. Among them, Voice over IP (VoIP) is one of the most important applications. In the early deployment of carrier-grade IP telephony, services were provided by gatekeepers [3], SIP proxy servers [4], or call agents that control media gateways [5, 6]. These mechanisms have been evolved into a more intelligent call control architecture called softswitch, which supports multi-protocols and provides interworking among heterogeneous networks. A softswitch is a cost-saving solution (in terms of capital and operation) that can flexibly enable new value-added telecommunications services and support the integration of web and multimedia services [7]. In 3G all-IP networks such as Universal Mobile Telecommunications System (UMTS) [8] the softswitch can be utilized as a 3G call agent or media gateway controller (MGC; see Figure 12.1 (6)). Unlike a traditional switch that includes hardwired-switching fabrics, the Softswitch (Figure 12.1 (6)) is a pure software-control mechanism without switching circuit hardware. The circuit hardware is implemented in separate media gateways (see Figure 12.1 (7)). Figure 12.1 illustrates the UMTS all-IP network architecture [1, 2], which consists of following components:

1. The radio access network consists of *User Equipments* (UEs; Figure 12.1 (1)) and *UMTS Terrestrial Radio Access Network* (UTRAN; Figure 12.1 (2)).
2. The GPRS network (Figure 12.1 (3)) consists of *serving GPRS support nodes* (SGSNs) and *gateway GPRS support nodes* (GGSNs) that provide mobility management and packet data services to mobile users.
3. The *home subscriber server* (HSS; Figure 12.1 (4)) is the master database containing all 3G user-related subscription information.
4. The *call state control junction* (CSCF; Figure 12.1 (5)) is a SIP server, which is responsible for call control.
5. The softswitch architecture consists of the *softswitch* or *media gateway controller* (Figure 12.1 (6)), *media gateway* (Figure 12.1 (7)) and *signaling gateway* (Figure 12.1 (8)).

In this section, we elaborate on the softswitch architecture (in Figure 12.1 (6)~(8)). Then based on this architecture, we describe the design and implementation of a softswitch we developed in Computer and Communication Research Laboratory (CCL), Industrial Technology Research Institute (ITRI).

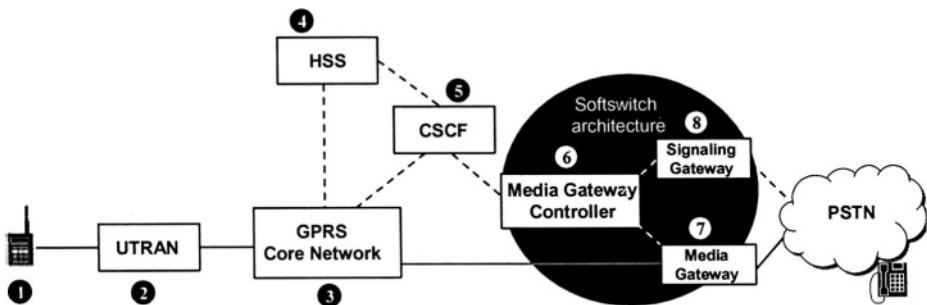


Figure 12.1. UMTS all-IP Network (dashed lines: signaling; solid lines: data and signaling)

1.1 ISC Reference Architecture

The International Softswitch Consortium (ISC) [9] is the premiere forum for the worldwide advancement of softswitch interoperability. This forum promotes Internet-based multimedia communications and applications. Figure 12.2 illustrates the reference softswitch architecture proposed by ISC. This architecture consists of four planes:

1. *Transport Plane* (see (1) in Figure 12.2) is used to transport packets across the IP network. These packets include call control signals and media data. The Transport Plane includes three domains:

- (i) *IP Transport Domain* consists of devices such as routers and switches that form the IP backbone. The QoS assured mechanisms (such as RSVP, Diffserv, MPLS, etc.) are also included in this domain.

- (ii) *Interworking Domain* consists of devices that interwork with other networks. For example, the trunking gateways (TG) and signaling gateways (SG) interwork with the PSTN and SS7 networks. The VoIP gateways interwork with other VoIP networks.
 - (iii) *Non-IP Access Domain* consists of devices that interwork with non-IP terminals and mobile networks. For example, access gateways (AG) and residential gateway interwork with non-IP terminals or phones; radio access network (RAN) AGs interwork with mobile network (GPRS/3G) devices; integrated access devices (IADs) and multimedia terminal adaptors (MTAs) interwork with broadband access (e.g. DSL/HFC) devices.
2. *Control & Signaling Plane* (see (2) in Figure 12.2) is the kernel of the softswitch architecture, which receives the signals from the Transport Plane and controls the flow of calls. The signaling protocols of this plane include traditional circuit-switched network protocols such as ISDN User Part (ISUP) [10] and Mobile Application Part (MAP) [11], and VoIP protocols such as H.323 [3], Media Gateway Control Protocol (MGCP) [5], Megaco [6] and Session Initiation Protocol (SIP) [4]. Softswitches reside in this plane.
3. *Service Plane* (see (3) in Figure 12.2) consists of application servers, which provide the service creation environment (SCE) and the service logic execution environments (SLEE). The SCE is a toolkit with graphic user interface (GUI), which allows users to develop services easily and quickly. The SLEE compiles service logics and communicates with the Control & Signaling Plane via standard Application Programming Interface (API; such as JAIN [12] and Parlay [13]) or protocols such as Intelligent Network Application Part (INAP) [14], Customized Application for Mobile Network Enhanced Logic (CAMEL) [15] and SIP. This plane also consists of media servers that provide conferencing and voice mail services.
4. *Management Plane* (see (4) in Figure 12.2) handles the management issues of the above three planes. The issues include service provisioning, network management, billing, etc.

The ISC reference architecture provides an interoperability view of softswitch, which is followed by the softswitch systems developed by Lucent [16], Siemens [17], Telcordia [18], and many other softswitch products and prototypes. In this chapter, we concentrate on the Control & Signaling Plane of the ISC Reference Architecture. Details of other planes are out of the scope of this chapter, and the reader is referred to [9].

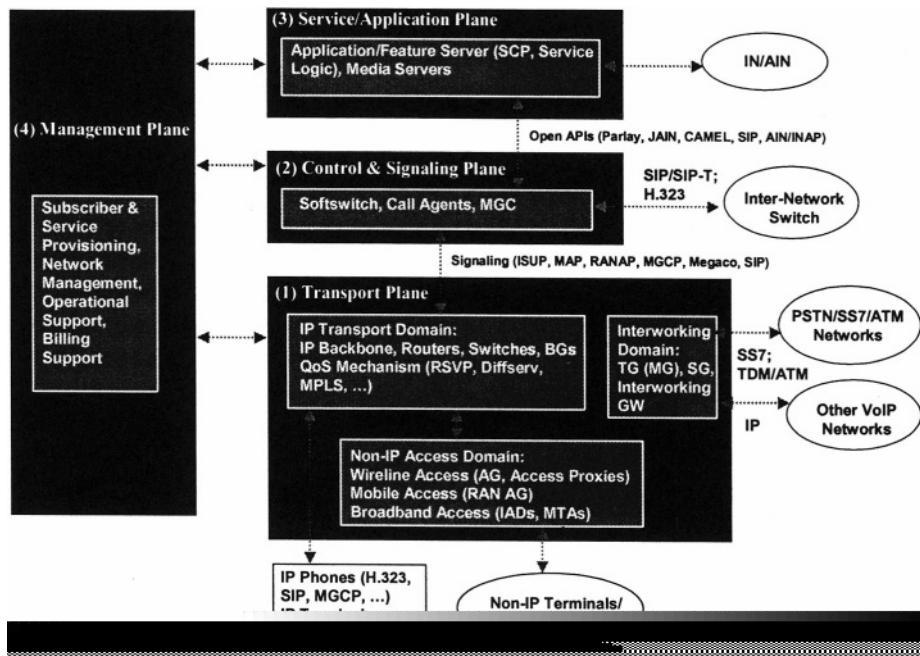


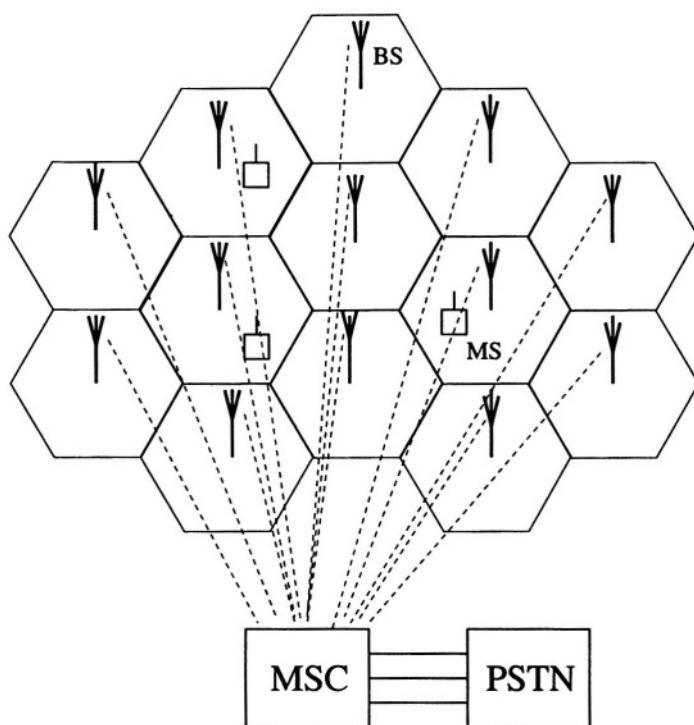
Figure 12.2. ISC Reference Architecture

1.2 CCL Softswitch Architecture

This section describes the softswitch modules and the design philosophy of our approach. Figure 12.3 illustrates the CCL Softswitch, which focuses on the Control & Signaling Plane. This architecture consists of the following modules:

1. *Protocol Adapter* (see (1) in Figure 12.3) provides mapping between the Transport Plane and the Call Server. The adapter converts the signals of various Transport-Plane protocols (such as ISUP, MGCP, Megaco, SIP and H.323) into a uniform format and delivers them to the Call Server through the Softswitch Service Provider Interface (SPI). With the adapter module, new protocols can be easily plugged into the system. Current implementation of the CCL Softswitch adapter module supports the following Transport-Plane protocols:

- (i) Off-net calls: Call setup between the PSTN and a VoIP system (including H.323, MGCP, Megaco and SIP)



BS: Base Station **MS:** Mobile Station

MSC: Mobile Switching Center

PSTN: Public Switched Telephone Network

Figure 16.1. The communication architecture for wireless cellular systems

eration wireless systems (such as the introduction of *gateway location register* or *GLR*), the fundamental operations in mobility management remains more or less the same. Hence, we will use the IS-41 standard as the baseline study here. In two-tier mobility management architecture shown in Figure 16.2, the HLR stores the user profiles of its registered mobiles, which contain the user profile information such as the mobile's identification number, the type of services subscribed, the quality of service (QoS) requirements and the current location information. The VLR stores the replications of user profiles and the temporary identification number for those mobiles which are currently visiting the associated RA. There are two basic operations in location management: *registration* and *location tracking*. The registration (also called *location update* in many situations) is the process that a mobile informs the system of its current location, the location tracking (also called *call delivery*) is the process that the system locates its mobile in order to deliver a call service to the mobile. When a user subscribes a service to a wireless network, a record of the mobile user's profile will be created in the system's HLR. Whenever and wherever the mobile user travels in the system's coverage, the mobile's location will be reported to the HLR (registration) according to some strategies. Then the location in HLR will be used to locate (*find*) the mobile. When the mobile visits a new area (called *registration area (RA)*), a temporary record will be created in the VLR of the visited system, and the VLR will then sends a registration message to the HLR. All the signaling messages are exchanged over the overlaying signaling network using signaling system 7 (SS7) standard.

To deliver a call service to a mobile from an originating switch (MSC), the calling mobile contacts its MSC, which then initiates a query to the caller's HLR to find where the callee is last reported (i.e., the VLR the callee was sighted). When the HLR receives this query, it will send a query to the VLR. The VLR, upon finding the mobile in its charging area, will return a routable address called the *temporary-location directory number (TLDN)* to the originating switch (MSC) through the HLR. Based on the TLDN, the originating switch (MSC) will set up the trunk (e.g., voice circuit with enough bandwidth for voice service) to the mobile. Thus, the call delivery process consists of two parts: *find*, the method to locate the mobile, and the trunk setup for the mobile using TLDN. The signaling traffic due to *find* and *registration (location update)* can be significant. Various schemes to reduce such traffic have been proposed. A very excellent survey on this topic is presented in [6]. There are mainly two approaches to reduce the signaling traffic. One is to manage the mobility databases physically or logically, the other is to control the location update frequency more intelligently. The funda-

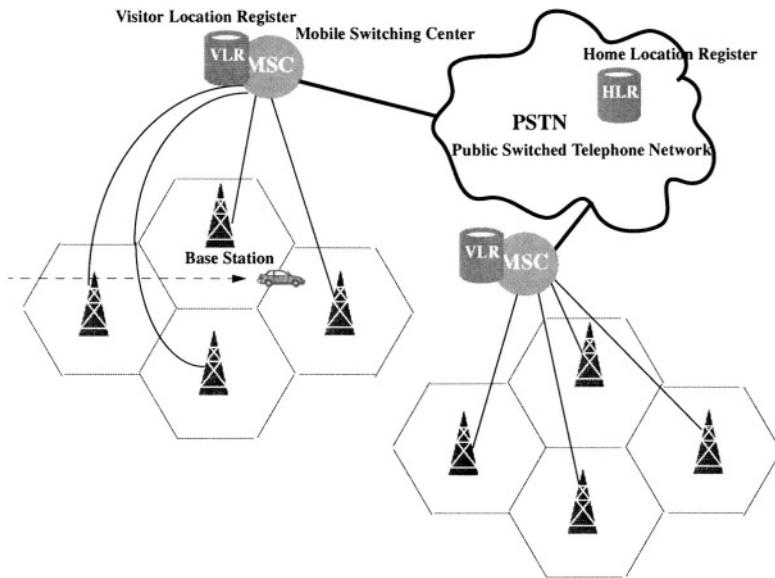


Figure 16.2. Two-tier location management architecture ([44])

mental idea in both approaches is to localize the signaling traffic. Here are a few representative schemes proposed in the past. In [35], the *location cache scheme* was proposed and shown significant improvement over IS-41 scheme when the frequency of the incoming calls is high with respect to the mobility. To reduce the signaling traffic from the mobile to the HLR, the *pointer forwarding scheme* ([34]) was proposed based on the observation that it may be better to setup a forwarding pointer from the previous VLR to avoid more expensive registration from the mobile. By storing location profile, registration traffic can also be reduced, this idea leads to the *alternative location algorithm* (ALA) and *two location algorithm* (TLA) ([42] and references therein). By observing that signaling cost can be significantly affected by the location database distribution, Ho and Akyildiz proposed the *local anchor scheme* to localize the registration traffic ([30]) and the dynamic hierarchical database architecture using *directory registers* ([29]). Recently, Ma and Fang ([47]) proposed a two-level pointer forwarding scheme to combine the advantages of pointer forwarding scheme and local anchor scheme. To maintain the freshness of the location information, active location update (also known as *autonomous registration*) was also suggested in the standard as an option ([16]). How often the active location update should be done is still an open question. Three schemes were investigated in the past ([7]): *timer-based*, *movement-based* and *distance-based*. It has been

shown that the most effective scheme among the three is distance-based: whenever a mobile moves away from the previously reported spot in certain distance, the mobile will carry out the location update. However, the implementation is complicated as most mobiles do not have GPS receivers equipped. The time-based scheme is simple to use, which has been used in most mobiles, however, it is not efficient in terms of signaling traffic, particularly for the mobiles with low mobility. It turns out that the movement-based (MB) scheme strikes the balance between the other two in both efficiency and implementation, which will be the one we investigate in this chapter. The problem is how to choose the movement threshold. The larger the threshold, the larger the uncertainty region of the mobile; while the smaller the threshold, the smaller the uncertainty region, but the more the signaling traffic. All above schemes are for location management. Another major part of signaling traffic comes from the last segment during the call delivery process: the paging. When the callee's VLR receives a query for the callee, a paging process is initiated. Depending on the information available and resource consideration, sequential paging, parallel paging or selective paging can be chosen ([56]).

To effectively evaluate a mobility management scheme, we have to deal with signaling traffic caused by both location update and call delivery. How to quantify the signaling traffic and the cost analysis become important. Signaling traffic cost relies on many factors such as the terminating call arrivals and the users' mobility. In the past, cost analysis of most mobility management schemes were carried out under the assumption that most time variables are exponentially distributed. For example, the time between two *served* calls, which we called *inter-service time* ([20]), was usually assumed to be exponentially distributed. Although the call arrivals terminating at a mobile, say, \mathcal{T} , may be approximately modeled by Poisson process, the inter-service time is not identical to the inter-arrival time due to the *busy-line effect* ([20]), i.e., some call arrivals for the mobile \mathcal{T} may not be connected because \mathcal{T} is serving another call, this is particularly true for data connections that tend to connect to the wireless systems longer than voice connection on average. Hence the served calls are in fact a "sampled" Poisson process, thus will be most likely not Poisson process. The reason we are interested in the inter-service times is that during the call connection of a mobile, no signaling traffic is necessary from this mobile since the network knows where the mobile is, only when the mobile is idle, does it need to make location update to inform the system. Thus, the location management is affected by the inter-service times rather than the inter-arrival times to the mobile. Moreover, due to the new trend of applications and user habits,

even the inter-arrival time for the terminating calls to a mobile may not be exponentially distributed anymore. Some adaptive or dynamic schemes for choosing some location management parameters depend on the explicit form of cost ([43]), whether such schemes are still effective or not when the inter-service times are not exponentially distributed is a question, there is no justification in the literature.

Recently, we have developed a new analytical approach to systematically analyze the performance of a few mobility management schemes under more realistic assumptions ([17, 18, 47]). In this chapter, we will present this analytical approach for the modeling and performance analysis for mobility management. We focus on the general analytical results for the signaling cost analysis for a few mobility management schemes we have studied in the past to illustrate our approach. Obviously, our results can be used to design the dynamic location management schemes.

This chapter is organized as follows. In the next section, we present the descriptions of a few known location management schemes we will study in this chapter. We then present a general framework for the evaluation of mobility management. The crucial analytical result on the probability distribution of the number of area boundary crossings are given in the fourth section. In the fifth section, we present the general analytical results for the signaling costs for the mobility management schemes we are interested in. We will conclude this chapter in the last section.

2. Mobility Management Schemes

There are many location management schemes in the literature, [6] presents a very comprehensive survey on all aspects of mobility management. In this chapter, we only concentrate on the schemes for which we have analytical results under fairly general assumptions. For some interesting schemes such as per-user caching scheme ([35]), location anchoring scheme ([30]), and location profile based scheme ([52]), we do not have the analytical results as yet, we will not include them in this section. More details can be found in [6].

2.1 IS-41 Scheme

To set the baseline comparison study, we first briefly go over the IS-41 scheme (or GSM MAP), the de facto standards for second generation wireless systems. We use the terminology used in [34]. An operation *move* means that a mobile user moves from one Registration Area (RA) (also called Location Area (LA)) to another while an operation *find* is the process to determine the RA a mobile user is currently visiting. The

move and *find* in second generation location management schemes (such as in IS-41 or GSM MAP) are called *basic move* and *basic find*. In the *basic move* operation, a mobile detects if it is in a new RA. If it is, it will send a registration message to the new VLR, the VLR will send a message to the HLR. The HLR will send a de-registration message to the old VLR, which will, upon receiving the de-registration message, send the cancellation confirmation message to the HLR. The HLR will also send a cancellation confirmation message to the new VLR. In the *basic find*, call to a mobile T is detected at a local switch. If the called party is in the same RA, the connection can be setup directly without querying the HLR. Otherwise, the local switch (VLR) queries the HLR for the callee, then HLR will query the callee's VLR. Upon receiving callee's location, the HLR will forward the location information to the caller's local switch, which will then establish the call connection with the callee. Periodic location update is optional for improving the efficiency of IS-41 and GSM MAP ([15, 16]).

2.2 Movement-based Mobility Management

There are many ways to improve the performance of the mobility management detailed in the IS-41. To minimize the signaling traffic due to the location update while keeping the location information fresh, we need to determine when we would need location update. In the current literature, three location update schemes were proposed and studied ([3, 7]): *distance-based location update*, *movement-based location update* and *time-based location update*. In distance-based locate update scheme, location update will be performed when a mobile terminal moves d cells away from the cell in which the previous location update was performed, where d is a distance threshold. In the movement-based location update scheme, a mobile terminal will carry out a location update whenever the mobile terminal completes d movements between cells (whenever the mobile moves from one cell to another, we count it as one move), where d is the movement threshold. In the time-based location update scheme, the mobile terminal will update its location every d time units, where d is the time threshold. It has been shown ([7]) that the distance-based location update scheme gives the best result in terms of signaling traffic, however, it may not be practical because a mobile terminal has to know its own position information in the network topology. The time-based location update scheme is the simplest to implement, however, unnecessary signaling traffic may result (imagine a terminal stationary for a long period may not need to do any update before it moves). The movement-based location update scheme seems to be the best choice

in terms of signaling traffic and implementation complexity. We will analyze the movement-based mobility management scheme here.

2.3 Pointer Forwarding Scheme (PFS)

The PFS modifies the *move* and *find* used in the IS-41 in the following fashion ([34]). When a mobile \mathcal{T} moves from one RA to another, it will inform its local switch (and VLR) at the new RA, which will then determine whether to invoke the *basic move* or the *forwarding move*. In the *forwarding move*, the new VLR exchanges messages with the old VLR to setup pointer from the old VLR to the new VLR, but does not involve the HLR. A subsequent call to the mobile \mathcal{T} from some other switches will invoke the *forwarding find* procedure to locate the mobile: queries the mobile's HLR as in the *basic find*, and obtains a “potentially outdated” pointer to the old VLR, which will then direct the *find* to the new VLR using the pointer to locate the mobile \mathcal{T} . To ensure that the time taken by the *forwarding find* is within the tolerable time limit, the length of the chain of the forwarding pointers must be limited. This can be done by setting up the threshold for chain length to be a number, say, K , i.e., whenever the mobile \mathcal{T} crosses K RA boundaries, it will register itself through the *basic move* (i.e., basic registration with HLR). In this way, the signaling traffic between the mobile and HLR can be curbed potentially.

2.4 Two-Level Pointer Forwarding Scheme

Two-level pointer forwarding scheme (TLPDS) is a generalization of PFS, which attempts to localize the location update signaling traffic and is consistent with the current hierarchical architecture of wireless system design ([47]). Instead of using one-level pointers in PFS, we activate a new layer of management entities, called *Mobility Agent (MA)*, the VLR which is in charge of multiple RAs, reflecting the regional activities of mobile users. The Two-Level Pointer Forwarding Scheme modifies the basic procedures used in IS-41 as follows. When a mobile moves from one RA to another, it informs the switch (and the VLR) at the new RA about the old RA. It also informs the new RA about the previous MA it was registered at. The switch at the new RA determines whether to invoke the *BasicMOVE* (update to the HLR) or the *TwoLevelFwdMOVE* (update to either the previous RA or the previous MA). In *TwoLevelFwdMOVE*, the new VLR exchanges messages with the old VLR or the old MA to set up a forwarding pointer from the old VLR to the new VLR. If a pointer is set up from the previous MA, the new VLR is selected as the current MA. The *TwoLevelFwdMOVE* procedures do not involve

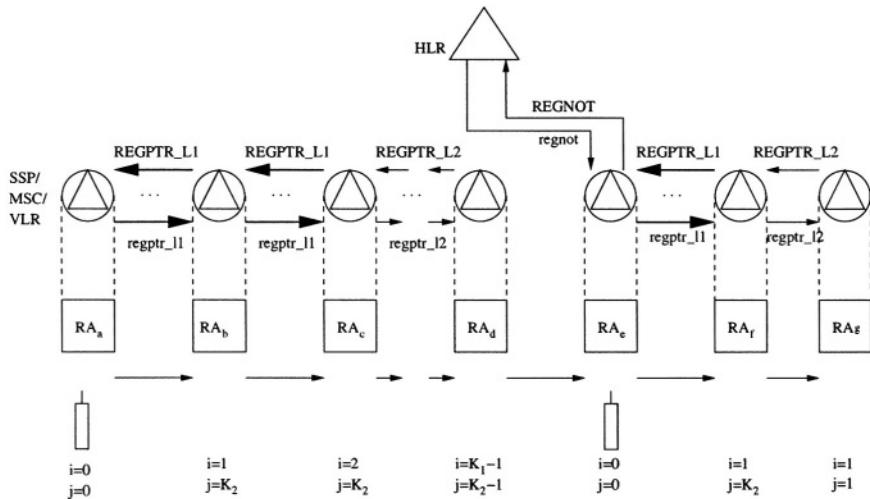


Figure 16.3. TLPDS location management architecture ([47])

the user's HLR. For example, Figure 16.3 shows a *Two-Level Forward MOVE* procedures with level-1 pointers chain threshold limited to 3. Assume that a user moves from RAa to RAg (these RAs are not necessary to be adjacent) and RAa is the user's MA. When the user leaves RAa but before enters RAb, the user informs the new VLRs and the level-2 pointers are built from the old VLR to the new VLR. When the user enters RAb, the chain threshold for level-1 pointer is reached, so RAb is selected as the user's new MA and a level-1 pointer is set up from the old MA to the new MA. At the same time, level-2 pointer chain is reset. The similar procedures are used at RAc. A level-1 pointer is set up from RAb to RAc, and the VLR in RAc is the new user's MA. As the user keeps moving. In RAe, the threshold of level-2 pointer chain is reached again, while this time the threshold of the level-1 pointer chain is reached too. Instead of exchanging information with the previous MA, the *BasicMOVE* is invoked. The HLR is updated with the user current location. The messages REGPTR.L1 and REGPTR.L2 are messages from the new VLR to the old VLR specifying that a level-1 or level-2 forwarding pointer is to be set up; messages regptr.l1 and regptr.l2 are the confirmations from the old VLR (or MA). In this figure, the VLRs in RAa, RAb, RAe and RAf are selected as the user's MAs.

To locate a mobile user, the *Find* procedure, called *TwolevelFwdFIND* procedure, is invoked for the subsequent calls to the mobile user from some other switches. The user's HLR is queried first as in the basic strategy, and a pointer to the user's potentially outdated MA is obtained.

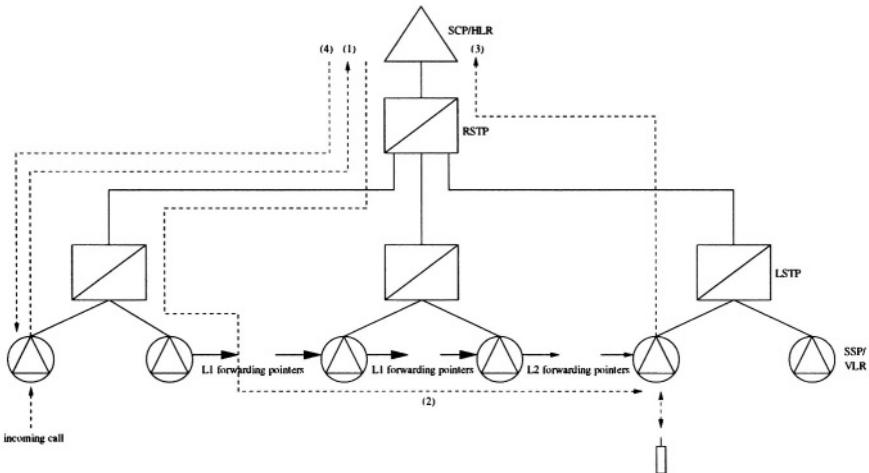


Figure 16.4. TLPDS Find procedure ([47])

The pointer chain is followed to find out the user's current location (see Figure 16.4). As we can see, in the two-level forwarding scheme, the chain length can be longer than that in the basic pointer forwarding scheme without increasing the *Find* penalty significantly. The previous study [34] shows that more saving can be obtained with longer chain. However, the pointer chain length is limited by the delay restriction requirement. By appropriately tuning the two thresholds in our schemes, we can mitigate the signaling cost without too much increase in the call setup delay.

2.5 Two Location Algorithm (TLA)

In the TLA ([42]), a mobile \mathcal{T} has a small built-in memory to store the addresses for the two most recently visited RA's. The record of \mathcal{T} in the HLR also has an extra field to store the two corresponding two locations. The first memory location stores the most recently visited RA. The TLA guarantees that the mobile \mathcal{T} is in one of the two locations. When the mobile \mathcal{T} joins the network, it registers with the network and stores the location in its memory and updates the HLR with its location. When the mobile \mathcal{T} moves to another registration area, it checks whether the RA is in the memory or not. If the new RA is not in the memory, the most recently visited (MRV) RA in the memory is moved to the second memory entry while the new RA is stored in the MRV position in the memory of \mathcal{T} . At the same time, a registration operation is performed to make the same modification in the HLR record. If the address for

the new RA is already in the memory, swaps the two locations in the memory of \mathcal{T} and no registration is needed and no action is taken in HLR record. Thus, in TLA, no registration is performed when a mobile moves back and forth in two locations. The consequence is that the location entries in the mobile and HLR may not be exactly the same: the MRV RA in HLR may not be the MRV RA in reality!

When a call arrives for the mobile \mathcal{T} , the two addresses are used to find the actual location of the mobile \mathcal{T} . The order of the addresses used to locate \mathcal{T} will affect the performance of the algorithm. If \mathcal{T} is located in the first try (i.e., a *location hit*), then the *find* cost is the same as the one in IS-41 scheme. Otherwise, the second try (due to the *location miss* in the first try) will find \mathcal{T} , which incurs additional cost. This additional cost has to be lower than the saved registration cost in order to make the TLA effective, which will demand a tradeoff analysis.

3. A General Framework for Performance Evaluation

Consider a wireless mobile network with cells of the same size. A mobile terminal visits a cell for a time interval which is generally distributed, then moves to the neighboring cell with equal probability (we are interested in this paper the homogeneous wireless networks in which all cells are statistically identical and all registration areas (RAs) are statistically identical). When a mobile is engaging a service, the network knows the location of the mobile, hence the basic idea to evaluate the performance of a mobility management scheme is to study the overall signaling traffic caused by location update and call delivery (*Find* or *Paging*) when the mobile is NOT engaging communication through the network, i.e., we only need to study the overall signaling traffic during the inter-service time. To carry tradeoff analysis, the signaling traffic has to be quantified, hence certain cost mapping may be necessary. Based on the cost structure for signaling traffic, performance comparison can be made possible. Moreover, some mobility management schemes have some tunable parameters, which need to be chosen to optimize the performance. By quantifying the signaling traffic, optimization problems can be formulated and solved analytically.

We take the movement-based location update scheme as one example to illustrate the framework for the cost analysis. Let d denote the movement threshold, i.e., a mobile terminal will perform a location update whenever the mobile terminal makes d movements (equal to the number of serving cell switching) after the last location update. When an incoming call to a mobile terminal arrives, the network initiates the

paging process to locate the called mobile terminal. Thus, both location update and terminal paging will incur signaling traffic, the location updates consume the uplink bandwidth and mobile terminal power, while the terminal paging mainly utilizes the downlink resource, hence the cost factors for both processes are different. When uplink signaling traffic is high or power consumption is a serious consideration, it may be better to use terminal paging instead. Different users or terminals may have different quality of service (QoS) requirement, location update and paging may be designed to treat them differently. All these factors should be considered and a more general cost function, which reflects these considerations, is desirable. A reasonable cost should consider two factors: location update signaling and paging signaling, which can be captured by the following general cost function:

$$\mathcal{C}(d) = \mathcal{C}_u(N_u(d), \lambda_u, q_s) + \mathcal{C}_p(N_p(d), \lambda_p, q_s) \quad (16.1)$$

where $N_u(d)$ and $N_p(d)$ are the average number of location updates and the average number of paging messages under the movement-based location update and a paging scheme with movement threshold d , respectively, during a typical period of inter-service time, λ_u is the signaling rate for location updates at a mobile switching center (MSC) and λ_p is the signaling rate for paging at an MSC, q_s indicates the QoS factor, \mathcal{C}_u and \mathcal{C}_p are two functions, reflecting the costs for location updates and paging. Depending on the choice of the two functions, we can obtain different mobility management schemes, particularly, we can obtain the optimal movement threshold d to minimize the total cost $\mathcal{C}(d)$, which gives us the best tradeoff scheme.

In the current literature, the total cost is chosen to be the linear combination of the location update and the paging signaling traffic, i.e., $\mathcal{C}_c(N_u(d), \lambda_u, q_s) = UN_u(d)$, $\mathcal{C}_p(N_p(d), \lambda_p, q_s) = PN_p(d)$, where U is the cost factor for location update and P is the cost factor for terminal paging ([4, 40, 42]). When the signaling traffic for location updates in an area is too high, some mobile terminals may be advised to lower their location updates, this can be done by choosing a function \mathcal{C}_u to contain a factor $1/(\lambda_{\max} - \lambda_u)$ where λ_{\max} is the maximum allowable signaling rate for location update. For example, if we choose $\mathcal{C}_u = U/(\lambda_{\max} - \lambda_u)$, when λ_u is approaching λ_{\max} , the update cost will become huge, the optimization will prefer to use paging more, leading to fewer location updates. How to choose the appropriate cost functions is a challenging problem to be addressed in the future.

For other schemes, the total cost will consist of two parts as well. The first part reflects the cost resulting from the location update, while the second part is contributed by the call delivery cost (*Find* operation). In

the movement-based mobility management scheme we discussed above, since the call delivery cost mainly comes from the terminal paging, hence we can use the terminal paging cost. Thus, the remaining task is to find the quantities we need in the cost function for each scheme.

4. Probability of the Number of Area Boundary Crossings

In order to carry out the performance analysis of mobility management schemes, we need to model some of the time variables appropriately and compute the probability distribution of the number of RA boundary crossings or cell boundary crossings. For example, in the mobility-based mobility management, we need to find the average number of location updates during the inter-service time, hence need to determine the distribution of cell boundary crossings. In other schemes (such as the PFS or TLPDS), we need to find the probability distribution of RA crossings to determine the cost for location update and call delivery. The results in this subsection have been presented in [20]. Although the results there are for RA crossings, they are valid for the cell crossings too, the only change to be made is to replace the RA residence time by the cell residence time. As we will show that this distribution plays a significant role in our analytical approach. For the simplicity of presentation, in this section, we use *an area* to denote either a registration area (RA) or a cell so that we can use the results to determine the probability of the number of RA boundary crossings and the probability of the number of cell boundary crossings as long as we use the corresponding residence time.

Assume that the incoming calls to a mobile, say, \mathcal{T} , form a Poisson process. The time the mobile stays in an *area* is called the *area residence time (ART)*. The time the mobile stays in a *registration area (RA)* is called the *RA residence time (RRT)*. The time the mobile stays in a cell is called the *cell residence time (CRT)*. Thus, ART can be either RRT or CRT for later applications. We assume that the ART is generally distributed with a non-lattice distribution (i.e., the probability distribution does not have discrete component). The time between the end of a call served and the start of the following call served by the mobile is called the *inter-service time*. It is possible that a call arrives while the previous call served is still in progress ([42]). In this case, the mobile \mathcal{T} cannot accept the new call (the caller senses busy tone in this case and may hang up). Thus, the inter-arrival (inter-call) times for the calls terminated at the mobile \mathcal{T} are different from the inter-service times. This phenomenon is called the *busy line* effect. We emphasize here that

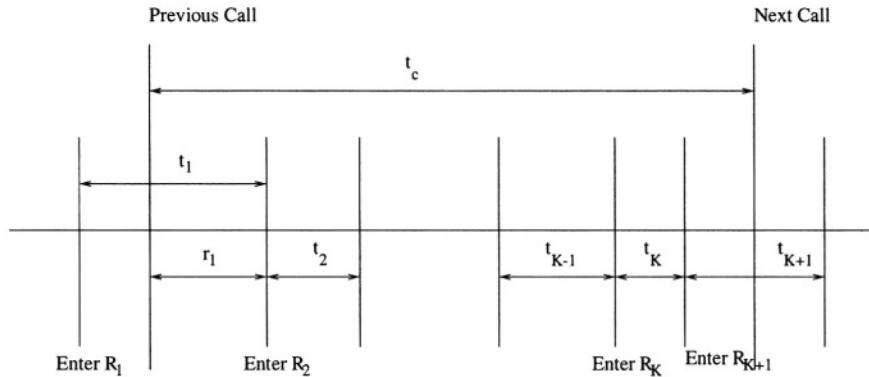


Figure 16.5. The time diagram for K area boundary crossings ([20])

for mobility management, it is the inter-service times, not the inter-arrival times of calls terminating at the mobile, that affects the location update cost, because when a mobile is engaging a call connection, the wireless network knows where the mobile \mathcal{T} is, hence the mobile does not need to carry out any location update during the call connection. Although the incoming calls form a Poisson process (i.e., the interarrival times are exponentially distributed), the inter-service times may not be exponentially distributed. Inter-service time bears the similarity to the inter-departure time of a queueing system with blocking, which has been shown to be non-exponential in general. In this chapter, we assume that the service time is negligibly shorter than the inter-service time (i.e., ignoring the service time) and derive the probability $\alpha(K)$ that a mobile moves across K areas between two served calls arriving to the mobile \mathcal{T} . In what follows, we will use the *area residence time (ART)* to denote either the RRT or the CRT, depending on the mobility management schemes we evaluate. By ignoring the busy line effect and the service time, assuming that the inter-service time is exponentially distributed, Lin ([42]) is able to derive an analytical formula for $\alpha(K)$, which has been subsequently used ([5, 4, 42, 55]) for tradeoff analysis for location update and paging. In this section, we assume that the inter-service times are generally distributed and derive an analytic expression for $\alpha(K)$.

Let t_1, t_2, \dots denote the area residence times and let r_1 denote the residual area residence time (i.e., the time interval between the time instant the mobile \mathcal{T} registers to the network and the time instant the mobile exits the first area). Let t_c denote the inter-service time between two consecutive served calls to a mobile \mathcal{T} . Figure 16.5 shows the time diagram for K area boundary crossings. Suppose that the mobile is in an

area R_1 when the previous call arrives and accepted by \mathcal{T} , it then moves K areas during the inter-service time, and \mathcal{T} resides in the j th area for a period t_j ($1 \leq j \leq K + 1$). In this chapter, we consider a homogeneous wireless mobile network, i.e., all areas (either registration areas or cells) in the network are statistically identical, hence ARTs t_1, t_2, \dots are independent and identically distributed (iid) with a general probability density function $f(t)$. We want to point out that the independence assumption is crucial for our derivation, it is not known yet in the current literature how all performance analysis in the wireless networks can be carried out analytically without this assumption. Let t_c be generally distributed with probability density function $f_c(t)$, and let $f_r(t)$ be the probability density function of r_1 . Let $f^*(s)$, $f_c^*(s)$ and $f_r^*(s)$ denote the Laplace-Stieltjes (L-S) transforms (or simply Laplace transforms) of $f(t)$, $f_c(t)$ and $f_r(t)$, respectively. Let $E[t_c] = 1/\lambda_c$ and $E[t_i] = 1/\lambda_m$. From the residual life theorem ([38]), we have

$$\begin{aligned} f_r(t) &= \lambda_m \int_t^\infty f(\tau) d\tau = \lambda_m [1 - F(t)], \\ f_r^*(s) &= \frac{\lambda_m}{s} [1 - f^*(s)], \end{aligned} \quad (16.2)$$

where $F(t)$ is the distribution function of $f(t)$. As a remark, the residual life theorem is valid only when we deal with the steady-state case, we can independently model the area residence time r_1 in the initiating area, i.e., we treat r_1 as a random variable which does not have any relationship to the ART t_i . The results we present here can be applied to this very general case. It is obvious that the probability $\alpha(K)$ is given by

$$\alpha(0) = \Pr[t_c \leq r_1], \quad K = 0, \quad (16.3)$$

$$\alpha(K) = \Pr[r_1 + t_2 + \dots + t_K < t_c \leq r_1 + t_2 + \dots + t_{K+1}], \quad K \geq 1 \quad (16.4)$$

Applying the inverse Laplace transform, we can compute $\alpha(0)$ from (16.3) as follows:

$$\begin{aligned} \alpha(0) &= \int_0^\infty \Pr(r_1 \geq t) f_c(t) dt = \int_0^\infty \frac{1}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \frac{1 - f_r^*(s)}{s} e^{st} ds f_c(t) dt \\ &= \frac{1}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \frac{1 - f_r^*(s)}{s} \int_0^\infty f_c(t) e^{st} dt ds \\ &= \frac{1}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \frac{1 - f_r^*(s)}{s} f_c^*(-s) ds \end{aligned} \quad (16.5)$$

where σ is a sufficiently small positive number which is appropriately chosen for the inverse Laplace transform.

For $K > 0$, $\alpha(K)$ is computed as follows. First, we need to compute $\Pr(r_1 + t_2 + \cdots + t_k \leq t_c)$ for any $k > 0$. Let $\xi = r_1 + t_2 + \cdots + t_k$. Let $f_\xi(t)$ and $f_\xi^*(s)$ be the probability density function and the Laplace transform of ξ . From the independence of r_1, t_2, t_3, \dots , we have

$$f_\xi^*(s) = E[e^{-s\xi}] = E[e^{-sr_1}] \prod_{i=2}^k E[e^{-st_i}] = f_r^*(s)(f^*(s))^{k-1}.$$

Thus the probability density function is given by

$$f_\xi(t) = \frac{1}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} f_r^*(s)(f^*(s))^{k-1} e^{st} ds.$$

Noticing that the Laplace transform of $\Pr(\xi \leq t)$ is $f_\xi^*(s)/s$, applying the inverse Laplace transform, we have

$$\begin{aligned} \Pr(r_1 + t_2 + \cdots + t_k \leq t_c) &= \int_0^\infty \Pr(\xi \leq t) f_c(t) dt \\ &= \int_0^\infty \frac{1}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \frac{f_r^*(s)[f^*(s)]^{k-1}}{s} e^{st} ds f_c(t) dt \\ &= \frac{1}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \frac{f_r^*(s)[f^*(s)]^{k-1}}{s} f_c^*(-s) ds. \end{aligned}$$

Taking this into (16.4), we obtain

$$\begin{aligned} \alpha(K) &= \Pr(t_c \geq r_1 + t_2 + \cdots + t_K) - \Pr(t_c \geq r_1 + t_2 + \cdots + t_{K+1}) \\ &= \frac{1}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \frac{f_r^*(s)[f^*(s)]^{K-1}[1 - f^*(s)]}{s} f_c^*(-s) ds \end{aligned} \tag{16.6}$$

where σ is a sufficiently small positive number. In summary, we obtain

THEOREM 16.1 *If the probability density function of the inter-service time has only finite possible isolated poles (which is the case when it has a rational Laplace transform), then the probability $\alpha(K)$ that a mobile moves across K areas (either RAs or cells) during the inter-service time is given by*

$$\begin{aligned} \alpha(0) &= \frac{1}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \frac{1 - f_r^*(s)}{s} f_c^*(-s) ds \\ \alpha(K) &= \frac{1}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \frac{f_r^*(s)[1 - f^*(s)][f^*(s)]^{K-1}}{s} f_c^*(-s) ds, \quad K > 0 \end{aligned} \tag{16.7}$$

where σ is a sufficiently small positive number and $f_r^*(s) = \lambda_m(1 - f^*(s))/s$.

If the inter-service time is exponentially distributed, then we can apply the Residue Theorem ([39]) to obtain the following simple result:

$$\begin{aligned}\alpha(0) &= 1 - \frac{1 - f^*(\lambda_c)}{\rho}, \\ \alpha(K) &= \frac{1}{\rho}[1 - f^*(\lambda_c)]^2[f^*(\lambda_c)]^{K-1}, \quad K > 0,\end{aligned}$$

where $\rho = \lambda_c/\lambda_m$ is the *call-to-mobility ratio*. This result has been obtained in [42] using a different approach.

5. Performance Evaluation

Signaling traffic in mobility management schemes incur from the operations of *move* (may cause location update) and *find* (may need message exchanges and terminal paging). Most performance evaluation for *move* and *find* were carried out under the assumption that some of the time variables are exponentially distributed. The conclusions drawn from such results may not be extended to the cases when such an assumption is not valid, and the adaptive schemes for choosing certain parameters (such as the threshold for the number of pointers or the threshold in movement-based mobility management scheme) may not be appropriate accordingly. There are not much works handling the non-exponential situations. Recently, we have developed a new approach which handles the non-exponential situations. In this section, we present our analytical results for signaling cost.

5.1 Cost Analysis for IS-41

As a baseline comparative study, we present the cost analysis for IS-41 first. We use the same notation as before, t_c denotes the inter-service time (the inter-arrival time for calls terminated at a mobile \mathcal{T} if we neglect the busy-line effect) with average $1/\lambda_c$ and t_i denotes the RA residence time with average $1/\lambda_m$. Let M and F denote the total cost for *basic moves* during the inter-service time and the total cost for *basic find*, respectively (i.e., the costs incurred in the IS-41 scheme). Since all location management schemes will go through the *move* and *find* whenever a terminating call to a mobile \mathcal{T} arrives, the inter-service time forms the fundamental regenerative period for cost analysis, thus we only need to consider the signaling traffic incurred during this period.

For the IS-41 scheme, whenever the mobile crosses a RA boundary, a registration is triggered. We assume that the unit cost for a basic

registration (i.e., *basic move*) is m . From Theorem 16.1, M will be equal to the product of m and the average number of registrations incurred during the inter-service time, given by ($f(t)$ and $f^*(s)$ are for the RA residence time)

$$\begin{aligned}
 M &= m \sum_{K=0}^{\infty} K \alpha(K) \\
 &= \frac{m}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \frac{f_r^*(s)[1-f^*(s)]}{s} \left(\sum_{K=1}^{\infty} K(f^*(s))^{K-1} \right) f_c^*(-s) ds \\
 &= \frac{m}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \frac{\lambda_m[1-f^*(-s)]^2}{s^2} \cdot \frac{1}{[1-f^*(-s)]^2} f_c^*(-s) ds \\
 &= \frac{m\lambda_m}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \frac{1}{s^2} f_c^*(-s) ds \\
 &= m\lambda_m \int_0^{\infty} f_c(t) \left(\frac{1}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \frac{1}{s^2} e^{st} ds \right) dt \\
 &= m\lambda_m \int_0^{\infty} f_c(t) t dt = \frac{m\lambda_m}{\lambda_c} = \frac{m}{\rho}
 \end{aligned}$$

where $\rho = \lambda_c/\lambda_m$, which is called the *call-to-mobility ratio*.

The *basic find* operation consists of two parts. The first part includes the interactions between the originating switch and the HLR while the second part includes the interactions between the HLR, the VLR, the MSC (terminating switch) and the mobile. Thus, $F \geq m$. The *basic find* will be the paging cost in one RA. Thus, we obtain the total cost for IS-41 during the inter-service time is

$$\mathcal{C}_{IS-41} = M + F = \frac{m}{\rho} + F. \quad (16.8)$$

{From this result, we observe that the total cost for IS-41 only depends on the CMR (i.e., the first moment of RA residence time and inter-arrival time of terminating calls).

5.2 Cost Analysis for Movement-based Mobility Management (MB)

The periodic location update (autonomous registration) is only an option in the IS-41 or the GSM MAP. The operation for such location update is simple, however, the scheme may not be efficient. The movement-based mobility management (MB) is a better choice in both the efficiency and simplicity. As we mentioned, the location update cost

relies on the average number of location updates during the inter-service time. In this section, we present the analytical results for this scheme.

5.2.1 Average Number of Location Updates. Let d be the threshold for the movement-based location update scheme, i.e., a mobile will make a location update every d crossings of cell boundary when the mobile does not engage in service. Obviously, the average number of location updates during the inter-service time will determine the locate update cost. So, we first want to find this quantity. We assume that $f(t)$ denote the probability density for the cell residence time (CRT) in this section, all other notation will remain the same as before. Then, the average number of location update during an inter-service time interval under movement-based location update scheme can be expressed as ([4])

$$N_u(d) = \sum_{i=1}^{\infty} i \sum_{k=id}^{(i+1)d-1} \alpha(k) \quad (16.9)$$

where $\alpha(k)$ in this subsection denotes the probability that a mobile crosses k cells during the inter-service time. In what follows in this subsection, we present the computation for $N_u(d)$. Let

$$S(n) = \sum_{k=1}^{n-1} \alpha(k),$$

then, from Theorem 16.1, we obtain

$$\begin{aligned} S(n) &= \sum_{k=1}^{n-1} \frac{1}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \frac{f_r^*(s)[1-f^*(s)][f^*(s)]^{k-1}}{s} f_c^*(-s) ds \\ &= \frac{1}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \frac{f_r^*(s)[1-f^*(s)]}{s} \left(\sum_{k=1}^{n-1} [f^*(s)]^{k-1} \right) f_c^*(-s) ds \\ &= \frac{1}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \frac{f_r^*(s)[1-(f^*(s))^{n-1}]}{s} f_c^*(-s) ds \end{aligned} \quad (16.10)$$

Moreover, we have

$$\begin{aligned} \sum_{k=1}^N S(kd) &= \sum_{k=1}^N \frac{1}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \frac{f_r^*(s)[1-(f^*(s))^{kd-1}]}{s} f_c^*(-s) ds \\ &= \frac{1}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \frac{f_r^*(s)}{s} \left\{ \sum_{k=1}^N [1-(f^*(s))^{kd-1}] \right\} f_c^*(-s) ds \\ &= \frac{1}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \frac{f_r^*(s)}{s} \left\{ N - \frac{(f^*(s))^{d-1}[1-(f^*(s))^{Nd}]}{1-(f^*(s))^d} \right\} f_c^*(-s) ds \end{aligned} \quad (16.11)$$

Thus, from (16.9), (16.10) and (16.11), after some mathematical manipulation (details can be found in [17]), we obtain

$$\begin{aligned}
 N_u(d) &= \sum_{i=1}^{\infty} i [S((i+1)d) - S(id)] \\
 &= \lim_{N \rightarrow \infty} \left\{ \sum_{i=1}^N i [S((i+1)d) - S(id)] \right\} \\
 &= \lim_{N \rightarrow \infty} \left\{ NS((N+1)d) - \sum_{i=1}^N S(id) \right\} \\
 &= \frac{1}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \frac{f_r^*(s)(f^*(s))^{d-1}}{s[1 - (f^*(s))^d]} f_c^*(-s) ds
 \end{aligned}$$

where $\sigma > 0$ is a sufficiently small positive number.

In summary, we obtain

THEOREM 16.2 *If the inter-service time $f_c^*(s)$ has a finite number of poles (such as a proper rational function) and the cell residence time (CRT) is generally distributed with probability density function $f(t)$ and L-S transform $f^*(s)$, then the average number of location updates $N_u(d)$ is given by*

$$N_u(d) = \frac{1}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \frac{f_r^*(s)(f^*(s))^{d-1}}{s[1 - (f^*(s))^d]} f_c^*(-s) ds \quad (16.12)$$

where $f_r^*(s)$ is the L-S transform of the probability density of the residual cell residence time. If $f_c^*(s)$ has a finite number of poles, which is the case when it is a rational function, then we have

$$N_u(d) = - \sum_{p \in \sigma_c} \text{Res}_{s=p} \frac{f_r^*(s)(f^*(s))^{d-1}}{s[1 - (f^*(s))^d]} f_c^*(-s), \quad (16.13)$$

where σ_c is the set of poles of $f_c^*(-s)$.

5.2.2 Tradeoff Cost Analysis: A Case Study. In this subsection, we present some results for the tradeoff analysis under movement based location update scheme and some paging scheme with the linear cost functional for illustration purpose.

Location Update Cost:

Let m denotes the unit cost for location update (i.e., the cost for a *basic move* as in IS-41), then total location update cost is given by

$$\mathcal{C}_u(d) = -m \sum_{p \in \sigma_c} \operatorname{Res}_{s=p} \frac{f_r^*(s)(f^*(s))^{d-1}}{s[1 - (f^*(s))^d]} f_c^*(-s), \quad (16.14)$$

If the inter-service time t_c is exponentially distributed with parameter λ_c , then $f_c^*(s) = \lambda_c/(s + \lambda_c)$, from (16.14), we can easily obtain

$$\mathcal{C}_u(d) = \frac{mf_r^*(\lambda_c)[f^*(\lambda_c)]^{d-1}}{1 - [f^*(\lambda_c)]^d}. \quad (16.15)$$

This result was obtained in [40] (noticing that $f_r^*(s) = \lambda_m(1 - f^*(s))/s$) via a different approach.

Paging Cost:

We consider the paging strategy used in [40] for our case study. Consider the hexagonal layout for the wireless network, all cells are statistically identical. According to the movement-based location update scheme, a mobile terminal moves at most d cells away from the previous position where it performs the last location update. Thus, a mobile terminal will surely be located in a cell which is less than d cell away from the previously reported position. If we page in the circular area with d cells as radius and with the previously reported position as the center, then we can definitely find the mobile terminal. Thus, this paging scheme is the most conservative among all paging. If we let P denote the unit cost for each paging in a cell, the maximum paging cost for this paging scheme is given by ([40])

$$\mathcal{C}_p(d) = P(1 + 3d(d - 1)), \quad (16.16)$$

Total Cost:

The unit costs (cost factors) m and P can be chosen to reflect the significance of the signaling (which may be significantly different from each other because they use different network resources). Given m , P and the movement threshold d , the total cost for location update and paging will be given by

$$\begin{aligned} \mathcal{C}_{MB}(d) &= \mathcal{C}_u(d) + \mathcal{C}_p(d) \\ &= -m \sum_{p \in \sigma_c} \operatorname{Res}_{s=p} \frac{f_r^*(s)(f^*(s))^{d-1}}{s[1 - (f^*(s))^d]} f_c^*(-s) + P(1 + 3d(d - 1)). \end{aligned}$$

(16.17)

We observe that as long as we find the probability distributions of cell residence time and inter-service time, we can find the total cost using 16.17. Many numerical results ([17]) have shown that the cost function $\mathcal{C}_{MB}(d)$ is a convex function of d . If this is true, then we can find the unique optimal threshold d^* to minimize the cost. Unfortunately, we have not proved the convexity of $\mathcal{C}(d)$ yet.

5.3 Cost Analysis for Pointer Forwarding

In this subsection, we evaluate the performance of pointer forwarding scheme (PFS). We need to quantify the signaling traffic incurring in this scheme. Let M' and F' denote the corresponding costs for the pointer forwarding scheme during the inter-service time, in which every K moves will trigger a new registration, where K is the maximum pointer chain length. In this subsection, a move means the crossing from one RA to another. Let S denote the cost of setting up a forwarding pointer between VLRs during a pointer forwarding *move* and let T denote the cost of traversing a forwarding pointer between VLRs during a pointer forwarding *find*. We first derive M' and F' .

Suppose that a mobile \mathcal{T} crosses i RA boundaries during the inter-service time, then there are $i - \lfloor i/K \rfloor$ pointer creations (every K moves require $K - 1$ pointer creations) and the HLR is updated $\lfloor i/K \rfloor$ times (with pointer forwarding, because the mobile \mathcal{T} registers every K th move). Here, we use $\lfloor x \rfloor$ to denote the floor function, i.e., the largest integer not exceeding x . Let $f(t)$ and $f^*(s)$ denote the probability density function and its corresponding L-S transform of the RA residence time (RRT). Let $\alpha(k)$ denote the probability that the mobile crosses k RAs during the inter-service time. Then, we obtain

$$\begin{aligned} M' &= \sum_{i=0}^{\infty} \left[\left(i - \left\lfloor \frac{i}{K} \right\rfloor \right) S + \left\lfloor \frac{i}{K} \right\rfloor m \right] \alpha(i) \\ &= S \sum_{i=0}^{\infty} i \alpha(i) + (m - S) \sum_{i=0}^{\infty} \left\lfloor \frac{i}{K} \right\rfloor \alpha(i) \\ &= S \sum_{i=0}^{\infty} i \alpha(i) + (m - S) \sum_{r=0}^{\infty} r \left(\sum_{i=rK}^{(r+1)K-1} \alpha(i) \right) \end{aligned} \quad (16.18)$$

Let

$$X(K) = \sum_{r=1}^{\infty} r \left(\sum_{i=rK}^{(r+1)K-1} \alpha(i) \right)$$

$$S(n) = \sum_{k=1}^{n-1} \alpha(k)$$

then, from Theorem 16.1 and following the same procedure we used in the last subsection, we obtain

$$S(n) = \frac{1}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \frac{f_r^*(s)[1 - (f^*(s))^{n-1}]}{s} f_c^*(-s) ds, \quad (16.19)$$

$$\sum_{i=1}^N s(iK) = \frac{1}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \frac{f_r^*(s)}{s} \left\{ N - \frac{(f^*(s))^{K-1}[1 - (f^*(s))^{NK}]}{1 - (f^*(s))^K} \right\} f_c^*(-s) ds \quad (16.20)$$

and

$$\begin{aligned} X(K) &= \sum_{r=1}^{\infty} r [S((r+1)K) - S(rK)] \\ &= \frac{1}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \frac{f_r^*(s)(f^*(s))^{K-1}}{s[1 - (f^*(s))^K]} f_c^*(-s) ds \end{aligned} \quad (16.21)$$

Noticing that $X(1) = \sum_{i=0}^{\infty} i\alpha(i)$, from (16.18), we obtain

$$\begin{aligned} M' &= SX(1) + (m - S)X(K) \\ &= \frac{1}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \left[\frac{Sf_r^*(s)}{s[1 - f^*(s)]} + \frac{(m - S)f_r^*(s)(f^*(s))^{K-1}}{s[1 - (f^*(s))^K]} \right] f_c^*(-s) ds \end{aligned} \quad (16.22)$$

Next, we derive F' . After the last *basic move* operation (if any), the mobile \mathcal{T} crosses $n = i - K\lfloor i/K \rfloor$ RA boundaries. Let $\Theta(n)$ denote the number of pointers to be tracked in order to find the mobile \mathcal{T} in the pointer forwarding *find* operation. If the mobile visits a RA more than once (i.e., a “loop” exists among n moves), then $\Theta(n)$ may not need to trace n pointers, thus, $\Theta(n) \leq n$. From this argument and applying Theorem 16.1, we obtain

$$\begin{aligned} F' &= \sum_{i=0}^{\infty} T\Theta(i - K\lfloor i \rfloor) \alpha(i) + F = T \sum_{r=0}^{\infty} \sum_{k=0}^{K-1} \Theta(k) \alpha(rK + k) + F \\ &= T \sum_{k=0}^{K-1} \Theta(k) \left(\sum_{r=0}^{\infty} \alpha(rK + k) \right) + F \\ &= T \sum_{k=0}^{K-1} \Theta(k) \frac{1}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \frac{f_r^*(s)[1 - f^*(s)](f^*(s))^{k-1}}{s[1 - (f^*(s))^K]} f_c^*(-s) ds + F \end{aligned}$$

$$= \frac{T}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \frac{f_r^*(s)[1-f^*(s)]}{s[1-(f^*(s))^K]} \left(\sum_{k=0}^{K-1} \Theta(k)(f^*(s))^{k-1} \right) f_c^*(-s) ds + F$$

In summary and applying the Residue Theorem ([39]), we finally arrive at

THEOREM 16.3 *If the inter-service time $f_c^*(s)$ has a finite number of poles (such as a proper rational function) and the RA residence time is generally distributed with probability density function $f(t)$ and L-S transform $f^*(s)$, then we have*

$$\begin{aligned} M' &= \frac{1}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \left[\frac{Sf_r^*(s)}{s[1-f^*(s)]} + \frac{(m-S)f_r^*(s)(f^*(s))^{K-1}}{s[1-(f^*(s))^K]} \right] f_c^*(-s) ds \\ &= - \sum_{p \in \sigma_c} \text{Res}_{s=p} \left[\frac{Sf_r^*(s)}{s[1-f^*(s)]} + \frac{(m-S)f_r^*(s)(f^*(s))^{K-1}}{s[1-(f^*(s))^K]} \right] f_c^*(-s) \\ F' &= F + \frac{T}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \frac{f_r^*(s)[1-f^*(s)]}{s[1-(f^*(s))^K]} \left(\sum_{k=0}^{K-1} \Theta(k)(f^*(s))^{k-1} \right) f_c^*(-s) ds \\ &= F - \sum_{p \in \sigma_c} \text{Res}_{s=p} \frac{f_r^*(s)[1-f^*(s)]}{s[1-(f^*(s))^K]} \left(\sum_{k=0}^{K-1} \Theta(k)(f^*(s))^{k-1} \right) f_c^*(-s) \end{aligned}$$

where σ_c denotes the set of poles of $f_c^*(-s)$ and $\text{Res}_{s=p}$ denotes the residue at the pole $s = p$.

If the inter-service time is exponentially distributed with parameter λ_c , then we have $f_c^*(-s) = \lambda_c / (-s + \lambda_c)$, from Theorem 16.3, we obtain

$$\begin{aligned} M' &= \frac{Sf_r^*(\lambda_c)}{1-f^*(\lambda_c)} + \frac{(m-S)f_r^*(\lambda_c)(f^*(\lambda_c))^{K-1}}{1-(f^*(\lambda_c))^K} \\ F' &= F + \frac{Tf_r^*(\lambda_c)[1-f^*(\lambda_c)]}{1-(f^*(\lambda_c))^K} \left(\sum_{k=0}^{K-1} \Theta(k)(f^*(\lambda_c))^{k-1} \right) \end{aligned}$$

which were obtained in [34] with a slight different form.

The worst case for F' would be when all pointers are traced, i.e., when $\Theta(n) = n$. In this case, we have the following result:

$$\begin{aligned} F' &= F + \frac{T}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \frac{f_r^*(s)[1-f^*(s)]}{s[1-(f^*(s))^K]} \left(\sum_{k=0}^{K-1} k(f^*(s))^{k-1} \right) f_c^*(-s) ds \\ &= F - T \sum_{p \in \sigma_c} \text{Res}_{s=p} \frac{f_r^*(s)[1-K(f^*(s))^{K-1} + (K-1)(f^*(s))^K]}{s[1-(f^*(s))^K][1-f^*(s)]} f_c^*(-s) \end{aligned}$$

Thus, the total cost for PFS during the inter-service time can be computed as follows:

$$\mathcal{C}_{PFS} = M' + F' = F + \frac{1}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} g(s) f_c^*(-s) ds = F - \sum_{p \in \sigma_c} \text{Res}_{s=p} g(s) f_c^*(-s) \quad (16.23)$$

where

$$\begin{aligned} g(s) &= \frac{S f_r^*(s)}{s[1 - f^*(s)]} + \frac{(m - S) f_r^*(s) (f^*(s))^{K-1}}{s[1 - (f^*(s))^K]} \\ &\quad + \frac{T f_r^*(s) [1 - f^*(s)]}{s[1 - (f^*(s))^K]} \left(\sum_{k=0}^{K-1} \Theta(k) \right). \end{aligned}$$

5.4 Cost Analysis for Two-Level Pointer Forwarding

Two-Level Pointer Forwarding (TLPFS) is a generalization of PFS in the sense that a new level of mobility agents (MA) is added between the home systems and visiting systems to regionalize (or localize) the signaling traffic. In [47], we derive the signaling cost under exponentially distributed inter-service time. Although we could obtain more general analytical results for this scheme under more general assumption in a similar way used for PFS, for simplicity, we only present the results in [47] and leave the derivation of more general results to the readers.

We use the same notation as before, we let t_c denote the inter-service time with average $1/\lambda_c$ and let $f(t)$ denote the probability density function of the RA residence time with average $1/\lambda_m$. The call-to-mobility ratio (CMR) is then given by $\rho = \lambda_c/\lambda_m$. We assume that the inter-service time is exponentially distributed. We define \mathcal{C}_{IS-41} and \mathcal{C}_{TLPFS} as the total costs of updating the location information (location update) and locating the user (location tracking) during the inter-service time in IS-41 and in the TLPFS, respectively. The following parameters used in the TLPFS:

m = the cost of a single invocation of *BasicMOVE*.

M = the total cost of all the *BasicMOVES* during the inter-service time.

F = the cost of a single *BasicFIND*.

M' = the expected cost of all *TwoLevelFwdMOVES* during the inter-service time.

F' = the average cost of the *TwoLevelFwdFIND*.

S_1 = the cost of setting up a forwarding pointer (level_1 pointer) between MAs during a *Two-LevelFwdMOVE*.

S_2 = the cost of setting up a forwarding pointer (level_2 pointer) between VLRs during a *Two-LevelFwdMOVE*.

T_1 = the cost of traversing a forwarding pointer (level_1 pointer) between MAs during a *Two-LevelFwdFIND*.

T_2 = the cost of traversing a forwarding pointer (level_2 pointer) between VLRs during a *Two-LevelFwdFIND*.

K_1 = the threshold of level_1 pointer chain.

K_2 = the threshold of level_2 pointer chain.

$\alpha(k)$ = the probability that there are k RA crossings during the inter-service time.

Then, we have

$$\begin{aligned}\mathcal{C}_{IS-41} &= M + F = m/\rho + F, \\ \mathcal{C}_{TLPFS} &= M' + F'.\end{aligned}$$

Thus, we need to derive formulas for M' and F' .

In the TLPFS, we observe that the HLR is updated only every $K_1 \cdot K_2$ moves (K_1 and K_2 are the level_1 and level_2 pointer chain length thresholds, respectively), while forwarding pointers are set up for all other moves. If a mobile user crosses i RA boundaries during the inter-service time, then the HLR is updated $\lfloor \frac{i}{K_1 K_2} \rfloor$ times, there are also $\lfloor \frac{i}{K_2} \rfloor - \lfloor \frac{i}{K_1 K_2} \rfloor$ level_1 pointer creations (every K_2 moves may require a level_1 pointer creation but sometimes the HLR is updated and level_1 pointer is not set up), and there are the level_2 pointers created for all the rest $i - \lfloor \frac{i}{K_2} \rfloor$ moves. Thus, we obtain

$$\begin{aligned}M' &= \sum_{i=0}^{\infty} \left\{ \left\lfloor \frac{i}{K_1 K_2} \right\rfloor m + \left(\left\lfloor \frac{i}{K_2} \right\rfloor - \left\lfloor \frac{i}{K_1 K_2} \right\rfloor \right) S_1 \right. \\ &\quad \left. + \left(i - \left\lfloor \frac{i}{K_2} \right\rfloor \right) S_2 \right\} \alpha(i) \\ &= \sum_{i=0}^{\infty} i S_2 \alpha(i) + \sum_{i=0}^{\infty} \left\lfloor \frac{i}{K_2} \right\rfloor (S_1 - S_2) \alpha(i) \\ &\quad + \sum_{i=0}^{\infty} \left\lfloor \frac{i}{K_1 K_2} \right\rfloor (m - S_1) \alpha(i).\end{aligned}\tag{16.24}$$

The cost F' is derived as follows. After the last *BasicMove* operations (if any), the mobile user traverses $\left\lfloor \frac{i - \left\lfloor \frac{i}{K_1 K_2} \right\rfloor K_1 K_2}{K_2} \right\rfloor$ level_1 pointers and $i - \left\lfloor \frac{i}{K_1 K_2} \right\rfloor K_1 K_2 - \left\lfloor \frac{i - \left\lfloor \frac{i}{K_1 K_2} \right\rfloor K_1 K_2}{K_2} \right\rfloor K_2$ level_2 pointers. Thus, we obtain

$$\begin{aligned}
 F' &= F + \sum_{i=0}^{\infty} \left\{ \left\lfloor \frac{i - \left\lfloor \frac{i}{K_1 K_2} \right\rfloor K_1 K_2}{K_2} \right\rfloor T_1 \right. \\
 &\quad \left. + \left(i - \left\lfloor \frac{i}{K_1 K_2} \right\rfloor K_1 K_2 - \left\lfloor \frac{i - \left\lfloor \frac{i}{K_1 K_2} \right\rfloor K_1 K_2}{K_2} \right\rfloor K_2 \right) T_2 \right\} \alpha(i) \\
 &= F + (T_1 - K_2 T_2) \sum_{i=0}^{\infty} \left\lfloor \frac{i - \left\lfloor \frac{i}{K_1 K_2} \right\rfloor K_1 K_2}{K_2} \right\rfloor \alpha(i) \\
 &\quad + T_2 \sum_{i=0}^{\infty} \left(i - \left\lfloor \frac{i}{K_1 K_2} \right\rfloor K_1 K_2 \right) \alpha(i). \tag{16.25}
 \end{aligned}$$

For simplicity, we let $g = f_m^*(\lambda_c)$. We observe that all summations in both equations (16.24) and (16.25) bear close similarities to those in subsection 16.5.3, hence applying a similar procedure, we can find the analytical results for M' and F' . The details can be found in [47]. In summary, we obtain

THEOREM 16.4 *If the inter-service time is exponentially distributed and the RA residence time is generally distributed with probability density function $f(t)$ and L-S transform $f^*(s)$, then the total signaling cost for TLPFS is given by*

$$C_{TLPFS} = M' + F'$$

where $g = f_m^*(\lambda_c)$ and

$$\begin{aligned}
 M' &= \frac{S_2}{\rho} + \frac{(1-g)g^{K_2-1}(S_1 - S_2)}{\rho(1-g^{K_2})} + \frac{(1-g)g^{K_1 K_2-1}(m - S_1)}{\rho(1-g^{K_1 K_2})} \\
 F' &= F + \frac{[1 - K_1 K_2 g^{K_1 K_2-1} + (K_1 K_2 - 1)g^{K_1 K_2}]T_2}{\rho(1-g^{K_1 K_2})} \\
 &\quad + \frac{(T_1 - K_2 T_2)(1-g)[g^{K_2} - K_1 g^{K_1 K_2} + (K_1 - 1)g^{(K_1+1)K_2}]}{\rho g(1-g^{K_1 K_2})(1-g^{K_2})}.
 \end{aligned}$$

5.5 Cost Analysis for TLA

In the Two Location Algorithm (TLA), if HLR has a location miss for a call termination, i.e., the two-location for the called mobile are

not the same as the one stored in the HLR, additional signaling traffic will be necessary to setup the call connection. Thus, the probability that the HLR has a location miss for a call setup, say, w , is important to capture the signaling traffic. In [42], this probability is derived under the assumption that the inter-service time is exponential. In this subsection, we first derive a more general analytical result to compute this quantity under the assumption that the inter-service time is generally distributed, then we present the cost analysis for the TLA.

*i*From the argument in [42], $(1 - \omega)$ is the probability that the HLR has the correct view of the latest visited RA when a call arrives (i.e., a location hit occurs and the *find* cost for TLA is the same as that for IS-41). A location hit occurs either when the mobile has not moved since last served call arrival, or when the last location update is followed by an even number of moves during the inter-service time, or when there are an even number of moves with no location update during the inter-service time. Here, in this subsection, a move means a boundary crossing from one RA to another. Let w_1 denote the probability that there is no move during the inter-service time, let w_2 denote the probability that the last served call is followed by an even number of moves without location update during the inter-service time, and let w_3 denote the probability that there are an even number of moves without location update during the inter-service time. Then, we have $1 - w = w_1 + w_2 + w_3$, i.e., $w = 1 - w_1 - w_2 - w_3$. Let $f(t)$ denote the probability density function for the RA residence time (RRT) and other related notations remain the same as before. Let $\alpha(k)$ denote again the probability that a mobile user crosses k RA boundaries during the inter-service time. For w_1 , we have

$$w_1 = \Pr(r_1 \leq t_c) = \alpha(0) \quad (16.26)$$

Let θ denote the probability of a move without location update. Let $w_2(K)$ denote the probability that the last registration followed by an even number of moves without location update during the inter-service time given that there are K moves (RA crossings) during the inter-service time. We can easily obtain that ([42])

$$w_2(K) = (1 - \theta) \sum_{i=0}^{\lfloor (K-1)/2 \rfloor} \theta^{2i} = \frac{1 - \theta^{2\lfloor (K-1)/2 \rfloor + 2}}{1 + \theta}, \quad K > 0$$

where $\lfloor x \rfloor$ indicates the floor function, i.e., the largest integer not exceeding x . Noticing that $w_2(2i+2) = w_2(2i+1)$ for $i \geq 0$, we have

$$w_2 = \sum_{K=1}^{\infty} w_2(K) \alpha(K) = \sum_{i=0}^{\infty} w_2(2i+1) [\alpha(2i+1) + \alpha(2i+2)]$$

$$= \frac{1}{1+\theta} \sum_{i=1}^{\infty} (1-\theta^{2i}) [\alpha(2i-1) + \alpha(2i)] \quad (16.27)$$

The probability w_3 can be computed as follows:

$$w_3 = \sum_{i=1}^{\infty} \theta^{2i} \alpha(2i).$$

Thus, applying Theorem 16.1 and after some mathematical manipulations similar to the techniques we used for PFS and TLPFS (the details can be found in [18]), we obtain

$$\begin{aligned} w &= 1 - w_1 - w_2 - w_3 \\ &= 1 - \alpha(0) - \frac{1}{1+\theta}(1-\alpha(0)) + \frac{1}{1+\theta} \sum_{i=1}^{\infty} \theta^{2i} - \frac{\theta}{1+\theta} \sum_{i=1}^{\infty} \theta^{2i} \alpha(2i) \\ &= \frac{\theta}{1+\theta} - \frac{\theta}{(1+\theta)2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \frac{1-f_r^*(s)}{s} f_c^*(-s) ds \\ &\quad + \frac{\theta^2}{(1+\theta)2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \frac{f_r^*(s)[1-f^*(s)]}{s[1-\theta^2 f^{*2}(s)]} f_c^*(-s) ds \\ &\quad - \frac{\theta^3}{(1+\theta)2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \frac{f_r^*(s)[1-f^*(s)]f^*(s)}{s[1-\theta^2 f^{*2}(s)]} f_c^*(-s) ds \\ &= \frac{1}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \frac{\theta f_r^*(s)}{s[1+\theta f^*(s)]} f_c^*(-s) ds \end{aligned} \quad (16.28)$$

Applying the Residue Theorem ([39]), we obtain

THEOREM 16.5 *If the inter-service time is distributed with rational Laplace transform $f_c^*(s)$ and the RA residence time (RRT) is generally distributed with probability density function $f(t)$ and with the L-S transform $f^*(s)$, the probability of a location miss in TLA is given by*

$$w = \frac{1}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \frac{\theta f_r^*(s)}{s[1+\theta f^*(s)]} f_c^*(-s) ds \quad (16.29)$$

where σ is a sufficiently small positive number and $f_r^*(s)$ is the L-S transform of the residual RRT. If σ_c denotes the set of poles of $f_c^*(-s)$, then we have

$$w = - \sum_{p \in \sigma_c} \text{Res}_{s=p} \frac{\theta f_r^*(s)}{s[1+\theta f^*(s)]} f_c^*(-s). \quad (16.30)$$

where $\text{Res}_{s=p}$ denotes the residue at the pole $s = p$.

If the inter-service time is exponentially distributed, which is the case studied in [42], then we have $f_c^*(s) = \lambda_c/(s + \lambda_c)$, hence from Theorem 16.5, we obtain

$$w = -\operatorname{Res}_{s=\lambda_c} \frac{\theta f_r^*(s)}{s[1 + \theta f^*(s)]} \cdot \frac{\lambda_c}{-s + \lambda_c} = \frac{\theta f_c^*(\lambda_c)}{1 + \theta f^*(\lambda_c)},$$

which is exactly the same as in [42] after some simplification of the result in [42].

Now, we are ready to carry out the cost analysis for TLA. We still assume that the cost for *basic move* is m and the cost for the *basic find* is F . The *basic find* consists of two parts: the first part is the message exchange from a mobile to the HLR, which is more or less the cost for the *basic move*; the second part is the message forwarding from the HLR to the callee plus the possible terminal paging. Hence, usually we have $F \geq m$ and the the cost for the second part in the *basic find* is $F - m$. Suppose that the mobile moves across K RAs during the inter-service time. The conditional probability $\Pr[I = i|K]$ that i *location update* operations are performed among the K moves has a Bernoulli distribution:

$$\Pr[I = i|K] = \binom{K}{i} \theta^{K-i} (1-\theta)^i$$

where θ is the probability that when a mobile \mathcal{T} moves, the new RA address is in the mobile's memory. Then the average number of location update during the inter-service time for TLA is given by

$$\begin{aligned} n_{TLA} &= \sum_{K=0}^{\infty} \sum_{i=0}^K i \Pr[I = i|K] \alpha(K) = \sum_{K=0}^{\infty} \left(\sum_{i=0}^K i \binom{K}{i} \theta^{K-i} (1-\theta)^i \right) \alpha(K) \\ &= \sum_{K=0}^{\infty} K(1-\theta) \alpha(K) = (1-\theta) X(1) \\ &= \frac{1-\theta}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \frac{f_r^*(s)}{s[1-f^*(s)]} f_c^*(-s) ds \\ &= -(1-\theta) \sum_{p \in \sigma_c} \operatorname{Res}_{s=p} \frac{f_r^*(s)}{s[1-f^*(s)]} f_c^*(-s) \end{aligned} \quad (16.31)$$

where we have used (16.21). Thus, the total cost for registration during inter-service time is $c_1 = mn_{TLA}$.

For the operations of the second part in the *basic find* for TLA, if we have a location hit (i.e., the location entries in the memories of both HLR and the mobile are identical), the *find* cost will be the same as in *basic find*; if there is a location miss, extra cost from HLR to the VLR

(second part of the *basic find* operation) will incur, thus the total cost for the *find operation* in TLA will be

$$c_2 = (1 - \omega) \cdot F + \omega[F + (F - m)] = F + (F - m)\omega.$$

In summary, the total signaling cost during the inter-service time for TLA is given by

$$\begin{aligned} C_{TLA} &= c_1 + c_2 = \delta n_{TLA} + F + (F - m)\omega \\ &= F + \frac{1}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \left[\frac{(1-\theta)m}{s[1-f^*(s)]} + \frac{(F-m)\theta}{s[1+\theta f^*(s)]} \right] f_r^*(s) f_c^*(-s) ds \\ &= F - \sum_{p \in \sigma_c} \text{Res}_{s=p} \left[\frac{(1-\theta)m}{s[1-f^*(s)]} + \frac{(F-m)\theta}{s[1+\theta f^*(s)]} \right] f_r^*(s) f_c^*(-s) \end{aligned} \quad (16.32)$$

where σ_c is the set of poles of $f_c^*(-s)$.

6. Some Interesting Special Cases

We notice that all general results presented so far can be easily applied when the inter-service time is distributed with rational Laplace transform. Since distributions with rational Laplace transforms are dense in the space of probability distributions, we can always use a distribution model with rational Laplace transform to approximate any given distribution to any desired accuracy. Some well-known distribution models used in queueing theory, such as Coxian model and phase-type distribution model, can be used to model both inter-service time and the area residence time (RRT or CRT), from which we can easily obtain the analytical results for the performance evaluation of mobility management schemes we investigated in this chapter. Recently, a few distribution models have been proposed to model various time variables in the wireless cellular networks, the hyper-Erlang model and the SOHYP (Sum of Hyper-Exponential) model are two of the important models (see [23] and references therein). They all belong to the phase-type distribution model, however, they are simple enough for analytical analysis and general enough to capture the statistics of the random time variables of interest. Due to the simplicity and generality of the hyper-Erlang model, we will concentrate on this model in this subsection.

The *hyper-Erlang* distribution (a more appropriate term may be *mixed-Erlang distribution*) has the following probability density function and Laplace transform:

$$f_{he}(t) = \sum_{i=1}^M p_i \frac{(m_i \eta_i)^{m_i} t^{m_i-1}}{(m_i - 1)!} e^{-m_i \eta_i t} \quad (t \geq 0),$$

$$f_{he}^*(s) = \sum_{i=1}^M p_i \left(\frac{m_i \eta_i}{s + m_i \eta_i} \right)^{m_i}, \quad (16.33)$$

where

$$p_i \geq 0, \quad \sum_{i=1}^M p_i = 1,$$

and M, m_1, m_2, \dots, m_M are nonnegative integers, $\eta_1, \eta_2, \dots, \eta_M$ are positive numbers.

It has been demonstrated ([23, 37]) that the hyper-Erlang distribution can be used to approximate the distribution of any nonnegative random variable. Many distribution models, such as the exponential model, the Erlang model and the hyper-exponential model, are special cases of the hyper-Erlang distribution. More importantly, the moments of a hyper-Erlang distribution can be easily obtained. If ξ is hyper-Erlang distributed as in equation (16.33), then its k th moment is given by

$$E[\xi^k] = (-1)^k f_{he}^{*(k)}(0) = \sum_{i=1}^M p_i \frac{(m_i + k - 1)!}{(m_i - 1)!} (m_i \eta_i)^{-k}.$$

The parameters p_i , m_i and η_i ($i = 1, 2, \dots, M$) can be found by fitting a number of moments from field data in practice.

Assuming now that the inter-service time is hyper-Erlang distributed with distribution given in equation (16.33) with

$$\lambda_c = \left(\sum_{i=1}^M p_i / \eta_i \right)^{-1},$$

then, applying the Residue Theorem ([39]), we can easily obtain

$$\begin{aligned} \mathcal{C}_{MB}(d) &= m \sum_{i=1}^M p_i \frac{(-1)^{m_i-1} (m_i \eta_i)^{m_i}}{(m_i - 1)!} d^{(m_i-1)}(m_i \eta_i) \\ &\quad + P(1 + 3d(d-1)) \\ \mathcal{C}_{PFS} &= F + \sum_{i=1}^M p_i \frac{(-1)^{m_i-1} (m_i \eta_i)^{m_i}}{(m_i - 1)!} g^{(m_i-1)}(m_i \eta_i) \\ \mathcal{C}_{TLA} &= F + \sum_{i=1}^M p_i \frac{(-1)^{m_i-1} (m_i \eta_i)^{m_i}}{(m_i - 1)!} h^{(m_i-1)}(m_i \eta_i) \end{aligned}$$

where

$$d(s) = \frac{f_r^*(s)(f^*(s))^{d-1}}{s[1 - (f^*(s))^d]},$$

$$\begin{aligned}
g(s) &= \frac{Sf_r^*(s)}{s[1-f^*(s)]} + \frac{(m-S)f_r^*(s)(f^*(s))^{K-1}}{s[1-(f^*(s))^K]} \\
&\quad + \frac{Tf_r^*[s][1-f^*(s)]}{s[1-(f^*(s))^K]} \left(\sum_{k=0}^{K-1} \Theta(k) \right), \\
h(s) &= \left[\frac{(1-\theta)m}{s[1-f^*(s)]} + \frac{(F-m)\theta}{s[1+\theta f^*(s)]} \right] f_r^*(s)
\end{aligned}$$

and $x^{(i)}(s)$ denotes the i th derivative of the function $x(s)$ at point s .

If the inter-service time is hyper-exponentially distributed with Laplace transform

$$f_c^*(s) = \sum_{i=1}^M p_i \frac{\eta_i}{s + \eta_i},$$

then we have

$$\begin{aligned}
\mathcal{C}_{MB}(d) &= m \sum_{i=1}^M p_i \eta_i d(\eta_i) + P(1 + 3d(d-1)) \\
\mathcal{C}_{PFS} &= F + \sum_{i=1}^M p_i \eta_i g(\eta_i) \\
\mathcal{C}_{TLA} &= F + \sum_{i=1}^M p_i \eta_i h(\eta_i)
\end{aligned}$$

If the inter-service time is exponentially distributed with Laplace transform

$$f_c^*(s) = \frac{\lambda_c}{s + \lambda_c},$$

then we have

$$\begin{aligned}
\mathcal{C}_{MB}(d) &= m \lambda_c d(\lambda_c) + P(1 + 3d(d-1)) \\
\mathcal{C}_{PFS} &= F + \lambda_c g(\lambda_c) \\
\mathcal{C}_{TLA} &= F + \lambda_c h(\lambda_c)
\end{aligned}$$

7. Conclusions

In this chapter, we have presented a new analytical approach to carrying out the performance evaluation for mobility management in wireless mobile networks. We focus on a few mobility management schemes we have investigated in the past few years and present the analytical results for the signaling traffic under very general assumption. These analytical results can be used to choose the design parameters in the mobility

management. Moreover, from these results, possible adaptive mobility management schemes can be developed based on the analytical results we present in this chapter. It is our hope that this chapter will inspire more applications of this novel approach to solve other problems in wireless networking.

References

- [1] A. Abutaleb and V. O. K. Li, "Paging strategy optimization in personal communication systems," *ACM Wireless Networks*, vol.3, pp.195-204, 1997.
- [2] A. Abutaleb and V. O. K. Li, "Location update optimization in personal communication systems," *ACM Wireless Networks*, vol.3, pp.205-216, 1997.
- [3] I.F. Akyildiz and J.S.M. Ho, "Dynamic mobile user location update for wireless PCS networks," *ACM Wireless Networks*, vol.1, pp. 187-196, 1995.
- [4] I.F. Akyildiz, J.S.M. Ho and Y.-B. Lin, "Movement-based location update and selective paging for PCS networks," *IEEE/ACM Trans. Networking*, vol.4, no.4, pp.629-638, 1996.
- [5] I.F. Akyildiz and W. Wang, "A dynamic location management scheme for next generation multi-tier PCS systems," To appear in *IEEE Journal on Selected Areas in Communications*.
- [6] I.F. Akyildiz, J. McNair, J.S.M. Ho, H. Uzunalioglu and W. Wang, "Mobility management in next-generation wireless systems," *Proc. of the IEEE*, vol.87, no.8, pp.1347-1384, August 1999.
- [7] A. Bar-Noy, I. Kessler and M. Sidi, "Mobile users: to update or not update?" *ACM Wireless Networks*, vol.1, no.2, pp. 175-186, July 1994.
- [8] R. Caceres and V. N. Padmanabhan, "Fast and scalable handoffs for wireless internetworks", *Proc. ACM Mobicom '96*, pp. 56-66, 1996.
- [9] A. T. Campbell, J. Gomez, S. Kim, C.-Y. Wan, Z. R. Turanyi and A. G. Valko, "Comparison of IP micromobility protocols," *IEEE Wireless Communications*, February 2002.

- [10] A. T. Campbell, J. Gomez, A. G. Valko, "An overview of cellular IP," *IEEE Wireless Communications and Networking Conference (WCNC'99)*, New Orleans, September 1999.
- [11] F.M. Chiussi, D. A. Khotimsky and Santosh Krishnan, "Mobility management in third-generation all-IP networks", *IEEE Communications Magazine*, pp. 124-135, September 2002.
- [12] D. C. Cox, "Wireless personal communications: what is it?" *IEEE Personal Comm. Mag.*, pp.20-35, April 1995.
- [13] D. R. Cox, *Renewal Theory*. John Wiley & Sons, New York, 1962.
- [14] S. Das, A. Misra, P. Agrawal and S. K. Das, "TeleMIP: Telecommunications-enhanced mobile IP architecture for fast intradomain mobility," *IEEE Personal Communications*, pp. 50-58, August 2000.
- [15] EIA/TIA, "Cellular radio-telecommunications intersystem operations," *EIA/TIA Technical Report IS-41 Revision B*, 1991.
- [16] ETSI, *Digital cellular telecommunications system (phase 2+): mobile application part (MAP) specification (GSM 09.02 version 7.51 Release)*, 1998.
- [17] Y. Fang, "Movement-based location management and tradeoff analysis for wireless mobile networks," Accepted for publication in *IEEE Transactions on Computers*.
- [18] Y. Fang, "General modeling and performance analysis for location management in wireless mobile networks," *IEEE Transactions on Computers*, **51**(10), 1169-1181, October 2002.
- [19] Y. Fang, I. Chlamtac and H. Fei, "Analytical results for optimal choice of location update interval for mobility database failure restoration in PCS networks," *IEEE Transactions on Parallel and Distributed Systems*, vol.11, no.6, pp.615-624, June 2000.
- [20] Y. Fang, I. Chlamtac and Y. B. Lin, "Portable movement modeling for PCS networks" *IEEE Trans. Veh. Tech.*, **49**(4), 1356-1363, July 2000.
- [21] Y. Fang, I. Chlamtac and H. Fei, "Analytical results for optimal choice of location update interval for mobility database failure restoration in PCS networks," *IEEE Transactions on Parallel and Distributed Systems*, Vol.11, No.6, pp.615-624, June 2000.

- [22] Y. Fang, I. Chlamtac and H. Fei, "Failure recovery of HLR mobility databases and parameter optimization for PCS networks," *Journal of Parallel and Distributed Computing*, Vol. 60, No. 4, pp. 431-450, April 2000.
- [23] Y. Fang and I. Chlamtac, "Teletraffic analysis and mobility modeling for PCS networks," *IEEE Transactions on Communications*, Vol. 47, No. 7, pp. 1062-1072, July 1999.
- [24] Y. Fang, I. Chlamtac and Y. B. Lin, "Billing strategies and performance analysis for PCS networks," *IEEE Transactions on Vehicular Technology*, Vol. 48, No. 2, pp. 638-651, 1999.
- [25] Y. Fang, I. Chlamtac and Y. B. Lin, "Channel occupancy times and handoff rate for mobile computing and PCS networks," *IEEE Transactions on Computers*, Vol. 47, No. 6, pp. 679-692, 1998.
- [26] Y. Fang, I. Chlamtac and Y. B. Lin, "Modeling PCS networks under general call holding times and cell residence time distributions," *IEEE/ACM Transactions on Networking*. Vol. 5, No. 6, pp. 893-906, December, 1997.
- [27] Y. Fang, I. Chlamtac and Y. B. Lin, "Call performance for a PCS networks," *IEEE Journal on Selected Areas in Communications*, Vol. 15, No. 8, pp. 1568-1581, October, 1997.
- [28] E. Gustafsson, A. Jonsson and C. Perkins , "Mobile IP regional registration," *Internet draft, draft-ietf-mobileip-reg-tunnel-04.txt*, work in progress, March 2001.
- [29] J.S.M. Ho and I.F. Akyildiz, "Dynamic hierarchical database architecture for location management in PCS networks," *IEEE/ACM Transactions on Networks*, 5(5), 646-660, October 1997.
- [30] J.S.M. Ho and I.F. Akyildiz, "Local anchor scheme for reducing signaling costs in personal communications networks," *IEEE/ACM Transactions on Networks*, 4(5), 709-725, October 1996.
- [31] J. S. M. Ho and I. F. Akyildiz, "Mobile user location update and paging under delay constraints," *ACM Wireless Networks*, vol. 1, no. 4, pp. 413-425, December 1995.
- [32] D. Hong and S. S. Rappaport, "Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures," *IEEE Trans. Veh. Tech.*, vol.35, no.3, pp.77-92, 198

- [33] B. Jabbari, "Teletraffic aspects of evolving and next-generation wireless communication networks," *IEEE Comm. Mag.*, pp.4-9, December 1994.
- [34] R. Jain and Y.B. Lin, "An auxiliary user location strategy employing forwarding pointers to reduce network impacts of PCS," *ACM Wireless Networks*, **1**(2), 197-210, 1995.
- [35] R. Jain, Y.B. Lin, C.N. Lo and S. Mohan, "A caching strategy to reduce network impacts of PCS," *IEEE Journal on Selected Areas in Communications*, **12**(8), 1434-1445, 1994.
- [36] I. Katzela and M. Naghshineh, "Channel assignment schemes for cellular mobile telecommunication systems: a comprehensive survey," *IEEE Personal Communications*. vol.3, no.3, pp.10-31, June 1996.
- [37] F. P. Kelly, *Reversibility and Stochastic Networks*, John Wiley & Sons, New York, 1979.
- [38] L. Kleinrock, *Queueing Systems: Theory, Volume I*, John Wiley & Sons, New York, 1975.
- [39] W. R. LePage, *Complex Variables and the Laplace Transform for Engineers*. Dover Publications, Inc., New York, 1980.
- [40] J. Li, H. Kameda and K. Li, "Optimal dynamic mobility management for PCS networks," *IEEE/ACM Transactions on Networking*, vol.8, no.3, pp.319-327, 2000.
- [41] W. Li and A. S. Alfa, "A PCS network with correlated arrival process and splitted-rating channels," *IEEE Journal on Selected Areas in Communications*, **17**(7), 1318-1325, 1999.
- [42] Y. B. Lin, "Reducing location update cost in a PCS network," *IEEE/ACM Transactions on Networking*, vol.5, no.1, pp.25-33, 1997.
- [43] Y.B. Lin, "Determining the user locations for personal communications networks," *IEEE Transactions on Vehicular Technology*, **43**, 466-473, 1994.
- [44] Y.B. Lin and I. Chlamtac, *Wireless and Mobile Network Architectures*, John Wiley and Sons, New York, 2001.
- [45] Y.B. Lin, Y.R. Haung, Y.K. Chen and I. Chlamtac, "Mobility management: from GPRS to UMTS,", *Wireless Communications and Mobile Computing (WCMC)*, vol.1, no.4, 339-359, 2001.

- [46] Y. B. Lin, S. Mohan and A. Noerpel “Queueing priority channel assignment strategies for handoff and initial access for a PCS network,” *IEEE Trans. Veh. Technol.*, vol.43, no.3, pp.704-712, 1994.
- [47] W. Ma and Y. Fang, “Two-level pointer forwarding strategy for location management in PCS networks,” *IEEE Transactions on Mobile Computing*, Vol.1, No.1, 32-45, Jan.-Mar. 2002.
- [48] S. Mohan and R. Jain, “Two user location strategies for personal communications services,” *IEEE Personal Communications*, pp.42-50, First quarter, 1994.
- [49] C. E. Perkins, “Mobile IP,” *IEEE Communications Magazine*, pp. 84-99, May 1997.
- [50] R. Ramjee, K. Varadhan, L. Salgarelli, S. R. Thuel, S.-Y. Wang and T. La Porta, “HAWAII: a domain-based approach for supporting mobility in wide-area wireless networks,” *IEEE/ACM Trans. of Networking*, vol. 10, no. 5, June 2002.
- [51] C. Rose and R. Yates, “Minimizing the average cost of paging under delay constraints,” *ACM Wireless Networks*, vol.1, no.2, pp.211-219, 1995.
- [52] S. Tabbane, “An alternative strategy for location tracking,” *IEEE J. on Selected Areas in Comm.* vol.13, no.5, pp.880-892, 1995.
- [53] W. Wang and I.F. Akyildiz, “A dynamic location management scheme for next generation multi-tier PCS systems,” *IEEE Transactions on Wireless Communications*, Vol.1., No.1 , pp. 178-190, January 2002.
- [54] W. Wang and I.F. Akyildiz, “A new signaling protocol for intersystem roaming in next-generation wireless systems,” *IEEE Journal on Selected Areas in Communications*, Vol. 19, No. 10, pp. 2040-2052, October 2001.
- [55] W. Wang and I.F. Akyildiz, “Intersystem location update and paging schemes for multtier wireless networks,” *Proc. of the ACM MobiCom’2000*, Boston, August 2000.
- [56] W. Wang, I.F. Akyildiz and G. Stuber, “Effective paging schemes with delay bounds as QoS constraints in wireless systems,” *ACM Wireless Networks*, No. 5, pp. 455-466, September 2001.

- [57] J. Xie and I. F. Akyildiz, "A distributed dynamic regional location management scheme for mobile IP," *Proc. IEEE INFOCOM 2002*, pp. 1069-1078, 2002.

Chapter #17

EFFICIENT INFORMATION ACQUISITION AND DISSEMINATION IN PERVASIVE COMPUTING SYSTEMS THROUGH CACHING¹

Mohan Kumar and Sajal K. Das

The University of Texas at Arlington

Abstract: Pervasive Computing applications require continual and autonomous availability of ‘what I want’ type of information acquisition and dissemination in a proactive yet unobtrusive way. Mobility and heterogeneity of pervasive environments make this problem even more challenging. Effective use of middleware techniques, such as *caching*, can overcome the dynamic nature of communication media and the limitations of resource-poor devices. In pervasive systems data is needed by users, devices, services and applications whereas caching mechanisms developed for mobile and distributed systems cater mainly to devices and in some special cases to users. Pervasive computing environments present entirely new set of challenges because of the fact that data may be acquired and disseminated at various stages within the system. Therefore, novel caching mechanisms are needed that take into account demand-fetched and prefetched (or pulled), as well as broadcast (or pushed) data. In addition, cache maintenance algorithms should consider such features as heterogeneity, mobility, interoperability, proactivity, and transparency that are unique to pervasive environments.

Key words: Pervasive computing, heterogeneity, mobility, data availability, and caching.

¹This research work was carried out under support from the Texas Advanced Research Program # 003656-0108-2001

1. INTRODUCTION

The essence of pervasive computing is the creation of smart environments saturated with computing and communication capability, yet gracefully integrated with human users [Wei91]. Pervasive computing encompasses many different technologies including mobile computing, wireless communications, middleware infrastructure, intelligence building, and wearable computers. Pervasive computing faces many challenges - heterogeneity, invisibility, mobility, and interoperability [Sat01]. In pervasive computing environments, hardware and software entities are expected to function autonomously, continually and correctly. Research in pervasive computing has been progressing at a rapid rate over the recent past in several leading universities and industries around the world. The most intriguing aspect of pervasive computing is its applications in almost all areas of human activity. The enabling technologies such as smart sensors (e.g., UC Berkeley Network Sensor Platform) and Radio Frequency ID (RFID) tags and the new generation of intelligent appliances, embedded processors, wearable computers, handheld computers and so on will continue to play a vital role in improving human quality of life through the advancement of pervasive computing applications [Kum03].

Making data available to users, applications, and services proactively is critical to pervasive computing applications. Current caching and content delivery mechanisms developed for mobile and distributed systems are inadequate for such applications. This is because existing caching mechanisms do not take into account the heterogeneity of the devices and the networking environments, and other unique features of pervasive systems.

This chapter presents in Section 2 an overview of the challenges and issues in pervasive computing for acquisition and dissemination of information in pervasive computing environments. The suitability and inadequacies of existing caching mechanisms for information delivery are discussed in Section 3. Then we propose in Section 4 some extensions to the existing caching mechanisms to make them more suitable for pervasive computing. The extensions include: profiling where in the profile of the user is stored and exploited for caching; data staging that involves staging of data on surrogate machines to decrease latency and save bandwidth; content adaptation of documents wherein the format of the documents are adapted to device profiles; and cache relocation to follow the roaming user. Section 5 summarizes the role of cache in pervasive computing and Section 6 presents our proposed architecture. Section 7 concludes the paper.

2. CHALLENGES OF PERVASIVE COMPUTING

This section explores the challenges and issues in pervasive computing environments, in which applications and users access information sources using a variety of mobile and static devices such as sensors, handheld computers, personal computers, servers, etc. The devices possess heterogeneous capabilities with respect to display size, communication mechanisms, memory, battery power, and computing capabilities. Applications and users attached to these devices connect to the external world via cellular radio networks, local area wireless networks, dial-up connections, or broadband connections. These communication systems offer different functionalities and heterogeneous characteristics with respect to bandwidth, delay, and latency. Also dealing with different degrees of smartness in different smart spaces is a unique challenge. According to Weiser [Wei91], pervasive computing is the complete disappearance of computing technology from a user's consciousness. Allowing access to information and services transparently and unobtrusively is one of the most important goals of pervasive computing.

The emerging area of pervasive computing incorporates research findings from many areas of computer science to meet the challenges posed by a myriad of applications. Today, we do have the necessary hardware and software infrastructure to foster the growth of this area. What is necessary is the *glue* to put together existing entities in order to provide meaningful and unobtrusive services to a variety of applications [Kum03a]. There exist excellent visionary papers, highlighting the issues and challenges of pervasive computing [Sat01, Ban00] – that include: invisibility, proactivity, interoperability, heterogeneity, mobility, intelligence, and security.

The development of computing tools hitherto has been in general, ‘reactive’ or ‘interactive’. On the other hand, ‘human in the loop’ has its limits, as the number of networked computers will surpass the number of users within a few years. Users of pervasive computing applications would want to receive ‘what I want’ kind of information and services in a transparent fashion. In a thought provoking paper [Ten00], Tenenhouse envisions that a majority of computing devices in future will be proactive. Proactivity can be provided by the effective use of overlay networks. In pervasive computing environments, users interact with various embedded or invisible computers. A proactive environment ensures proactive acquisition and dissemination of context-aware information to users and applications.

Effective caching of information at various stages can enhance proactivity and improve its availability transparently.

Today's computing world is replete with numerous types of devices, operating systems, and networks. Cooperation and collaboration among various devices and software entities is necessary for pervasive computing. At the same time, the overheads introduced by the adaptation software should be minimal and scalable. While it is almost unthinkable (in today's environment) to have homogeneous devices and software, it is however, possible to build software bridges across various entities to ensure interoperability. The limitations of resource-poor hardware can be overcome by exploiting the concepts of agents and services. A major challenge is to develop effective mechanisms, such as caching at various stages in order to mask the uneven conditions and develop portable and lightweight application software.

Mobile computing devices have limited resources (CPU, memory, bandwidth, energy), are likely to be disconnected, and are required to react (transparently) to frequent changes in the environment. Mobile users desire *anytime anywhere* access to information while on the move. Typically, a wireless interface is required for communication among the mobile devices. The interface can be a wireless LAN, a cellular network, a satellite network, or a combination thereof. Techniques developed for routing, multicasting, caching, and data access in mobile environments need to be extended carefully to pervasive environments consisting of numerous invisible devices, anonymous users, and ubiquitous services. Development of effective information delivery mechanisms to mask the heterogeneous wireless networks and mobility effects is a significant challenge. Provisioning uniform services regardless of location is also a vital component of pervasive computing. The challenge here is to provide location-aware information availability in an adaptive fashion, in a form that is most appropriate to the location as well as the situation under consideration [Che02].

Mobility arises in several forms in pervasive environments – such as mobility of devices, mobility of users and mobility of software agents. Because users, devices, and agents change context, tasks like querying, scheduling and caching management become increasingly complex in pervasive computing systems.

To create smart spaces we need an intelligent, situation-aware system augmented by middleware support. The challenge is to incorporate middleware tools into smart environments. In a crisis management situation, for example, the remote sensing system should have abilities to infer the user's intention and

current situation to inform of the possible danger prior to the occurrence of the crisis. Smart home projects are being carried out in many universities and research laboratories [Cur99, Pin03]. Smart homes allow the occupants to interact with the home equipment and appliances by exploiting knowledge gained through user behavior. Current smart homes are not yet advanced enough to allow the user to take the (virtual) home with him/her. Dynamic, context-aware information can be pushed into the user's mobile device prior to leaving the smart home. For example, the Sentient computing environment [Cur99] is localized and very limited in terms of functionality. With the middleware services, the same environment can be used for pervasive health care as envisioned in [Had02].

Several methods for knowledge representation in intelligent spaces have been investigated. Provisioning proactive, context-aware or situation-oriented response from the infrastructure to the users' needs is a challenge. As smart spaces grow in sophistication, the intensity of interactions between a user's personal computing space and his/her surroundings increases. This has severe bandwidth, energy, and distraction implications for a wireless mobile user. The presence of multiple users will further complicate this problem. Scalability of applications is a serious issue on many fronts – number of users, physical spread of the client and variety of devices. Additionally, the application data is likely to have real-time component requiring specific delivery deadlines.

3. CURRENT CACHING MECHANISMS

A key requirement of pervasive computing is the ability of the applications to be aware of the characteristics which are part of the context for the application running on devices. For example, if the multimedia streaming application can get information about the lack/availability of wireless bandwidth it may choose to scale down/up the fidelity of the output. Some caching mechanisms do incorporate mobility issues and wireless communication channel characteristics in identifying data items for caching.

Consistency of a mobile client's cached items in disconnection-prone environments has been a subject of investigation [Cao02]. Some current schemes [Kah01] use *invalidation report* (IR) in which an IR is periodically broadcast by the server and clients use IRs to validate their cached items. There exist several problems with regard to IR-based schemes: i) the client or device may be offline or disconnected when the IR arrives; ii) the source should be queried to obtain the

updated data; and iii) the frequency of IR reports may be very high with respect to frequency of reference at certain user sites.

Some mechanisms such as the one by Tuah et al.[Tua02] decide on the data to be cached based on next access probability. Data items are cached based on probability of access in the future or popularity. Caching in low bandwidth networks increases data availability and reduces latency. User and device profiles can help determine the utility values of cacheable data. However, current caching mechanisms do not take into account the device profiles while caching data.

In the remainder of this section we discuss cache coherence and cache replacement mechanisms that have been proposed for mobile and distributed systems.

3.1 Cache Coherence

A cache coherence strategy usually executes in two phases - invalidation and update. Most applications in mobile environments will generate more read operations than write operations. Further-more a mobile client usually can accept a slightly out-dated data in return for faster data retrieval.

In the *refresh time method* [Yue00] of invalidation the estimation of the refresh time for an object depends on its update probability. If an object is updated frequently, its refresh time will be shorter. In this method, the refresh time for an object is updated dynamically whenever the object is sent to a client from the server. This approach does not require a client to be always connected in order to invalidate an object. Furthermore, if an object is never accessed again, it will never be refreshed even after its refresh time expires.

In *asynchronous invalidation* [Cao02] reports are broadcast by the server only when some data changes not periodically. Each mobile host maintains its own home location cache (HLC) to deal with the problem of disconnections. The HLC is maintained at a designated home mobile switching station (MSS). It has an entry for each data item cached by the mobile host and needs to maintain only the time-stamp at which the data item was last invalidated. At the cost of this extra memory overhead for maintaining an HLC, the mobile host can continue to use its cache even after

prolonged periods of disconnection from the network. A mobile host alternates between active mode and sleep mode. In the sleep mode, a mobile client is unable to receive any invalidation reports sent to it by the HLC. The time stamp is used to decide as to which invalidations to retransmit to the mobile host. When the mobile host wakes up after a sleep, it sends a probe message (along with the time stamp) to its HLC. In response to this probe message, the HLC sends it an invalidation report. Thus a mobile host can determine which data items got changed while it was disconnected. A mobile host defers all queries, which it receives after waking up until it has received the invalidation report from its HLC.

3.2 Cache Replacement Policies

Replacement policies adopt the access probabilities of database items as an indicator for the necessity of replacing a cached item. For each object a *replacement score* indicating the prediction of its access probability is estimated. The higher the score the higher is the estimated access probability and the lower is its chance of being replaced. The most straightforward way to compute the score for each object is by measuring the mean inter-operation arrival duration. The cached object with the highest mean arrival duration is replaced in the mean scheme.

The mean scheme probably does not adapt well to changes in access patterns since every single trace from the beginning of the access history remains in effect. A better approach is to use a window for the statistical measures. Each object is associated with a window of size W storing the access time of W most recent operations. The cached object with the highest mean arrival duration within the window is replaced. This is known as the *window scheme* [Yue00], whose effectiveness depends on the window size. A problem with the window scheme is the amount of storage needed to maintain the window values.

To avoid the need of a moving window and to adapt quickly to changes in access patterns, the exponentially weighted moving average [EWMA] scheme assigns weights to each arrival/duration. In EWMA recent durations have higher weights and the weights tail off as the durations get longer. The replacement score is the EWMA of arrival durations [Su02]. A parameter to the EWMA is the weight, α which ranges from 0 to 1. The current duration receives a weight of 1, the previous duration receives a weight of α , the next previous duration receives a weight of α^2 and so on.

In traditional cache replacement policies, access probability is considered as the most important factor that affects cache performance. A probability-based policy is to replace the data with the least access probability. In location-dependent services, besides access probability there are two other factors, namely data distance and valid scope area that should be considered in cache replacement. Data distance refers to the distance between the current location of a mobile client and the valid scope of a data value. In a location-dependent data service, the server responds to a query with the suitable value of the data item according to the client's current location. As such, when the valid scope of a data value is far away from the client's current location, this data will have a lower chance to become usable again since it will take some time before the client enters the valid scope area again and the data is useless before the user reaches the valid scope area. In this respect, we should favor ejecting the *farthest* data when replacement takes place. Valid scope area refers to the geometric area of the valid scope of a data value. For location-dependent data, valid scope areas can somehow reflect the access probabilities for different data values. That is, the larger the valid scope area of the data, the higher the probability that the client requests this data. This is because, generally, the client has a higher chance of being in large regions than small regions. Thus, a good cache replacement policy should also take this factor into consideration. Based on the above analysis, a promising cache replacement policy should choose as its victim, data with low access probability, a small valid scope area and a long data distance.

4. EXTENDING CURRENT CACHING MECHANISMS FOR PERVASIVE COMPUTING

As seen in the previous section the existing caching mechanisms in use today fall short of the requirements for pervasive computing. This section investigates techniques to make cache management strategies more suitable to meet the challenges of pervasive computing. Pervasive environments may comprise devices with varying communication, display and computing capabilities; and software agents representing disconnected users/devices. One of the challenges in pervasive computing is to adapt the environment to the user's needs by predicting the user's location and needs. In the following, we investigate various caching mechanisms and assess their suitability of pervasive environments. In the following, we discuss various caching strategies that can be adapted to pervasive environments.

4.1 Profiling

Data recharging is a service that aims to provide analogous functionality for recharging the device–resident data cache [Che01]. The goal of data recharging is to deliver data in a flexible and geographically independent manner. That is, users transparently recharge their devices with context-aware data by connecting at any point in an available infrastructure such as the Internet using available communication media. The data needed on a particular device is highly dependent on the user of that device and the tasks he/she is performing. Thus, the choice of data that must be sent to a device in order to recharge it is dependent on the semantics of the applications. The data recharging infrastructure maintains user profiles to facilitate transfer of user-specific data.

A user profile contains two types of information – the types of data that are of interest to the user and the user priorities. These profiles can play a very important role in caching as they could be used while caching data and also while choosing victims to eject from the cache.

4.2 Hierarchical Caching – Content Adaptation

In pervasive computing, users access information sources using a wide variety of mobile devices featuring heterogeneous capabilities with respect to size, memory, battery power and computing capability. The key to meeting the demands in such a heterogeneous environment is the adaptation of the representation of documents to the capabilities of the device. This is called content adaptation. This adaptation may include: format transcoding (e.g. XML, to WML, JPEG to WBMP) scaling of images as well as video and audio streams, media conversion and omission or substitution of document parts. Web caching exploits sharing and temporal locality in the access patterns of clients to satisfy requests by cached copies of the required objects. By terminating client requests at local caching proxies, the network traffic on the Internet as well as the server load is reduced whereas the response time is improved. Furthermore, the robustness of web services is enhanced as a client may retrieve a document from the proxy cache if the server is unavailable.

There are three different approaches for performing the adaptations - content server-side adaptation, proxy-based, and adaptation paths. In server-side adaptations, adapted documents are supplied to the clients by the server either by

on-demand dynamic adaptation or by having a repository of pre-adapted documents. With proxy-based adaptation the documents are served by the servers in a generic representation (e.g. XML-based). The adaptation is performed by intermediary proxies. Those proxies are usually placed close to the clients. The concept of adaptation paths has been introduced by the UC Berkeley's Ninja project [??]. As opposed to server-side adaptation, proxy based adaptation allows the caches to be placed between the information sources and adaptation proxies.

4.3 Data Staging

Data Staging is a technique that improves the performance of distributed file systems running on small, storage-limited pervasive computing devices by opportunistically caching data on nearby surrogates[Fli02]. The file system tries to reduce access time by caching files on the client machine, but space limitations prevent it from caching all but a portion of the files that the user might potentially read. This surrogate plays the role of a second level cache for a mobile client. By proactively staging data on the surrogate, cache misses from a nearby mobile client can be serviced at low latency.

5. ROLE OF CACHE IN PERVASIVE COMPUTING

Pervasive computing aims at seamlessly integrating computing and communication with our environment so as to make our day-to-day activities the principal focus rather than the computing or communication devices. As mobile and computing devices become more pervasive, the nature of interaction between users and computers is evolving. Computing applications need to become increasingly autonomous and invisible by placing greater reliance on context-awareness in order to increase user transparency. Moreover due to the current popular emphasis on *anytime anywhere* computing applications must cope with the highly dynamic environments in which resources such as network connectivity, bandwidth, security and software services vary over time and therefore applications behavior and functionality have to adapt to currently available resources and constraints. Thus we see that pervasive computing although taking shape today imposes a number of challenges. We investigate how a cache meets some of these challenges.

The data from the server has to be adapted and formatted before it can be pushed to the device. A combination of effective caching and middleware agents can be used to meet this challenge. A software agent associated with the

cache could be employed to adapt data from the server to meet the media handling capabilities of the device. This transformation of the data may include format transcoding, scaling of images as well as audio and video streams, media conversion and also omission or substitution of document parts. By doing this, the server can serve documents requested in a generic representation and these documents could be formatted /adapted by agents in the cache according to the capabilities of the device.

The second challenge is invisibility, i.e., how to make the interaction between the environment and computing devices unobtrusive. The user must be able to obtain just what he wants without being aware of the computing environment. Frequently accessed data should be made available anywhere to the clients through effective caching. Caching frequently accessed data in the device could improve the response time and hide the latency due to scarce bandwidth. The cache could be used to provide the user with exactly the information he requires by constantly replenishing the cache with user profiles. Thus caching makes the interaction between the environment and the user least obtrusive, by pushing only data that is important to the user in the current context.

When the user moves from one location to another, his requirements change, and the data required may not have been already cached. This will result in fetching relevant data items from the server, thus increasing the latency. This problem can be alleviated if user location prediction methods are employed. Based on this location prediction the cache could prefetch items relevant to the user's next location before the user moves there. Thus when the user moves to the next location, the data required by him in that location are already cached, thereby reducing access latency for these items.

One of the very important requirements of an effective pervasive system is proactivity. The environment should be able to sense the needs of the user and deliver information proactively. Caching and profiling algorithms can be integrated to ensure availability of 'what I want' kind of information in a pervasive computing environment.

6. ARCHITECTURE FOR PERVASING CACHING

The architecture of a caching mechanism for pervasive environments is shown in Figure 1. The main components of such a system are the cache on the receiving entity and the cache on the server or network. The receiving entity can be a user's device, user's software agent, a service or an application. A cache can

be associated with each of these entities. The information documents are cached based on their utility values. The algorithms for determining utility values, the optimization strategy, and incorporation of user/system inputs comprise the software components of the system. At any point of time, the cache comprises traditionally cached, prefetched and push-cached documents. The utility algorithm updates the utility values of i) cached documents ii) remote documents to be prefetched and iii) documents in the next broadcast cycle.

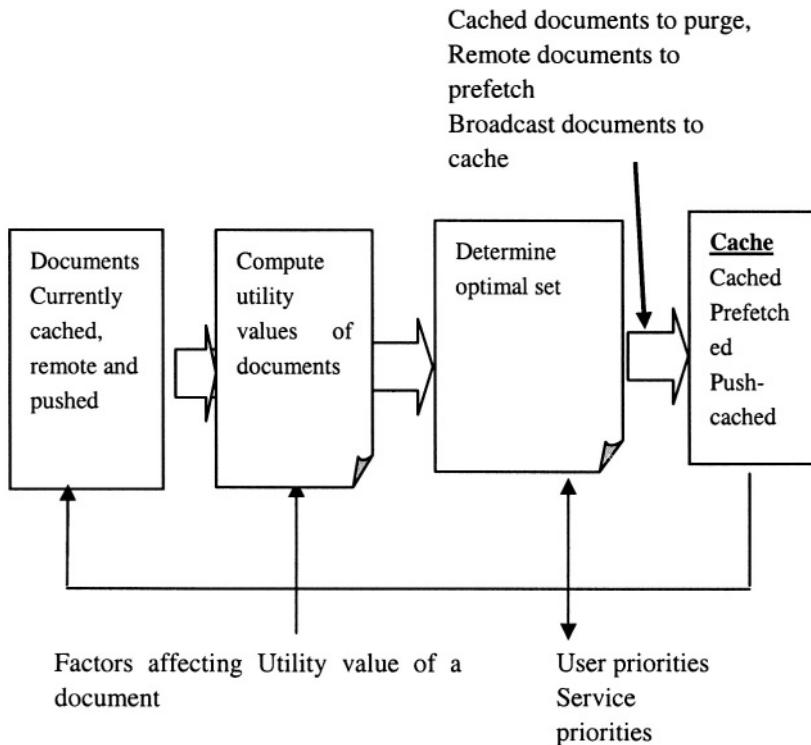


Figure 1 Architecture of a Caching Scheme for Pervasive Computing

6.1 Utility Functions

Each document D_i ($0 \leq i \leq n$) is represented by a tuple $D_i < s_i, u_i >$, where s_i is the document size and u_i is the utility of the document. A document's utility ($0 \leq u_i \leq 1$), varies dynamically based on various parameters and whether the document is cached, broadcast or to be prefetched. The factors pertinent to a document are shown in Table 1 below. The utility of each document is determined by normalizing the effect of several factors by employing a rule

similar to those discussed in Greedy-Dual caching strategies [Lau02]. In fact, developing such a rule for content delivery in wireless Internet is one of the main challenges of this project. Furthermore, the computation of utility values can be tuned to user/server requirements by applying weights to various factors in order to provide flexibility.

Table 1. Factors affecting utility value of a document

| | Description | Cache | Prefetch | Push |
|----------------------------|--|-------|----------|------|
| <i>Cost</i> | <i>The communication cost of transmission</i> | ✓ | ✓ | |
| <i>Mobility factor</i> | <i>The rate at which MH moves from node to node</i> | ✓ | ✓ | ✓ |
| <i>Hit ratio</i> | <i>The number of hits/the number of requests</i> | ✓ | | |
| <i>Delay saving ratio</i> | <i>Time to transmit once/time to transmit many times</i> | ✓ | ✓ | ✓ |
| <i>Access frequency</i> | <i>Number of times a document is accessed</i> | ✓ | | |
| <i>Dependence</i> | <i>The dependence relation with documents already in cache</i> | ✓ | ✓ | ✓ |
| <i>Bandwidth</i> | <i>Bandwidth available</i> | | ✓ | |
| <i>Client profile</i> | <i>Utility of the document based on profile</i> | ✓ | ✓ | |
| <i>Server profile</i> | <i>Popularity value for a broadcast document</i> | | | ✓ |
| <i>Device profile</i> | <i>Capabilities at the receiving device</i> | ✓ | ✓ | ✓ |
| <i>Broadcast frequency</i> | <i>The frequency at which a document is broadcast</i> | | | ✓ |
| <i>Update frequency</i> | <i>The frequency at which a document is updated</i> | ✓ | ✓ | ✓ |
| <i>Battery power</i> | <i>Battery power left on device</i> | ✓ | ✓ | |

The utility of a remote document D_i to be prefetched can be high if the wireless link between a device and the infrastructure has high bandwidth whereas u_i is low if the wireless link has low bandwidth, unavailable or congested. Likewise, the utility of a cached document decreases every time it is not requested. The mobility of the MH plays a significant role in determining the utility of a document. Thus each cached document's utility is updated continuously and the

utility of a remote document is updated whenever it is in contention for a spot in the cache.

6.2 Optimal Set of Documents

The next important step is to determine the optimal set of documents to cache, given document sizes and utilities. The objective is to maximize utility while at the same time make optimal use of the cache space. First, the weight w_i , of the object in the cache is calculated by using $w_i = (u_i \times C/s_i)$, where C is the cache capacity, u_i is the utility and s_i is the size of the document respectively. Then the documents are sorted in order of their weights. In a traditional system, with neither prefetch nor broadcast, documents can be cached based on their weights. However, the situation with pervasive environments is quite complex when one or more of the following events happen: i) wireless bandwidth is available, ii) a broadcast document is received, or iii) an extraordinary event has taken place. On occurrence of such an event, utilities and hence the weights of these documents change. In addition, user-defined and/or server-defined priorities to cache documents including quality of service (QoS) parameters can be incorporated into the architecture. The set of documents to be cached is determined by employing optimization techniques such as greedy algorithms and dynamic programming.

Utility values are updated for all cached documents having high access probability, and documents related to cached documents. Broadcast documents that are in the optimal set are cached and also the procedure for prefetching new documents in the optimal set which are not currently cached is initiated. Once prefetch documents arrive, cached documents that are not in the optimal set are purged to make room. Likewise an optimization algorithm is executed on every server or network node in order to maintain an optimal set of documents. Utility values are computed for all cached documents as well as for those in the prefetch lists. The next broadcast sequence is determined and then requests for transfer of remote documents are sent out on the Internet. Each of these algorithms should ideally execute in $O(n)$ time, where n is the number of documents considered.

6.3 Cache consistency

Assuming successful caching of required data documents, we have to deal with the cache consistency issues. Cached documents may be updated on the infrastructure, when that happens, there will be multiple copies of the same data document. Such copies may be on connected or disconnected

mobile devices. Therefore we should have special mechanisms to handle such situations that are uncommon in traditional systems. Stale copies of cached documents on mobile devices must be invalidated before they are accessed by the mobile client.

An important criterion for caching is the rate at which a particular data document is updated. If a data document is likely to be updated very frequently, then it's not prudent to cache it. Several researchers have investigated mechanisms to deliver invalidation reports efficiently to mobile clients. Update frequencies of documents should be considered in determining the utility values.

There can be two types of cache consistency maintenance algorithms for mobile computing environments: stateless and stateful. In the stateless approach, the server is unaware of the user's content. Stateless approaches use simple database management and they are less expensive to maintain. Even though, such approaches would suffice for many routine data dissemination applications. On the other hand, stateful approaches require large databases and complex maintenance algorithms. The Scalable Asynchronous Cache Consistency Scheme (SACCS) [Wan03] inherits positive features of both approaches. The SACCS mechanism employs flag bits to indicate data validity at the server and users' ends to minimize cost of consistency maintenance algorithms. An improvisation of SACCS can be used for pervasive environments to incorporate unique features of pervasive environments.

6.4 Multimedia delivery

A continuous or streaming media system allows large multimedia objects to be played out at clients before the entire object is delivered. There are many classes of continuous media (CM) applications relevant to pervasive computing. These classes include multi-party conferencing, peer-to-peer networks, online gaming, multimedia delivery and collaboration. Before CM data is rendered on an end-system, it typically undergoes capturing, encoding, storage, and delivery. There has been much success in using web caches to alleviate network-related bottlenecks. Caches are incrementally deployable solutions that require no support from the network, servers or even clients. As CM usage becomes more prevalent in pervasive systems, it is worthwhile to deploy caching that accommodates special characteristics of CM and related protocols. CM caching research is still in its infancy. Caching can potentially improve system throughput and increase data

availability. It is necessary to extend the existing schemes to address wireless multimedia issues such as delay, jitter and bandwidth availability in order to provide a high available, adaptive and consistent system. In [Lau02], a GD-based scheme considered traditional caching for continuous media delivery in wireless networks. For pervasive computing applications it would be necessary to include prefetch and broadcast documents and address heterogeneity and scalability issues.

7. CONCLUSIONS

Pervasive computing technologies and associated software are being employed to facilitate such applications as telemedicine, education, space endeavors, manufacturing, crisis management, transportation, and defense. In pervasive computing environments, hardware and software entities are expected to function autonomously, continually and correctly. Information acquisition and dissemination by the hardware and software entities in a pervasive computing scenario is a challenging problem. However, existing caching and content delivery mechanisms are inadequate to meet the demands on pervasive environments. This chapter presents an investigation of existing mobile and distributed caching strategies that can be improvised for pervasive computing applications. Furthermore, an architecture for facilitating information acquisition and dissemination using caching (comprising cached, prefetched and broadcast documents) strategies is presented.

References

- [Ban00] G. Banavar, J. Beck and E. Gluzberg, "Challenges: An application model for pervasive computing," in Proceedings of 6th Annual International Conference on Mobile Computing and Networking (MOBICOM 2000), pp.266-274, 2000, Boston MA, USA.
- [Cao02] Guohong Cao, On Improving the Performance of Cache Invalidation in Mobile Environments, Mobile Networks and Applications, Volume 7, Issue 4, August 2002.
- [Cha01] B.Y. Chan, A. Si, and H. V. Leong, A Framework for Cache Management for Mobile Databases: Design and Evaluation Distributed and Parallel Databases 10 (1):23-57, July 2001.
- [Che01] Cherniack, M., Franklin, M.J., Zdonik, S., Expressing user profiles for Data Recharging , IEEE Personal Communications, Volume: 8 Issue: 4 , Aug 2001 Page(s): 32 -38.

- [Che02] Liang Cheng and Ivan Marsic, Piecewise Network Awareness Service for Wireless Mobile Pervasive Computing, *Mobile Networks and Applications*, Volume 7 , Issue 4, August 2002.
- [Cur99] R. Curwen, A. Hopper, P. Steggles and A. Ward, "Sentient Computing," Technical Report 1999.13 (video), AT&T Laboratories Cambridge, 24a Trumpington Street, Cambridge CB2 1QA, England.
- [Fli02] Jason Flinn, Shafeeq Sinnamohideen and M. Satyanarayanan , Data Staging on Untrusted Surrogates, Intel Research, Pittsburgh IRP-TR-02-02 May 2002
- [Gup01] Gupta, S.K.S. ,Wang-Chien Lee , Purakayastha, A., Srimani, P.K , An overview of Pervasive Computing *IEEE Personal Communications*, Volume: 8 Issue: 4, Aug. 2001, Page(s): 8 -9.
- [Had02] Stathes Hadjiefthymiades, Vicky Matthaiou, Lazaros Merakos, On Supporting the WWW in Wireless Communications Through Mobile Agents , *Mobile Networks and Applications*, Volume 7 , Issue 4, August 2002.
- [Kah01] A. Kahol, S. Khurana, S.K.S. Gupta, and P. K. Srimani, A Strategy to Manage Cache Consistency in a Disconnected Distributed Environment, *IEEE Trans. on Parallel and Distributed Systems*, Vol. 12, No. 7, July 2001, page(s): 686-700.
- [Kum03]M Kumar, Pervasive Computing Education, *Distributed Systems Online*, July 2003, Vol. 4, No.7.
- [Kum03a] M Kumar, B Shirazi, S K Das, B Sung, D Levine, and M Singhal, Pervasive Information Communities Organization PICO: A Middleware Framework for Pervasive Computing *IEEE Pervasive Computing*, July-September 2003, pgs. 10-17.
- [Lau02] WHO Lau, M Kumar, and S. Venkatesh, A Cooperative Cache Architecture in Support of Caching Multimedia Objects in MANETs, Atlanta, September 2002.
- [Sat01] Satyanarayanan, M., Pervasive Computing: Vision and Challenges, *IEEE Personal Communications*, August 2001, Pgs. 10- 17.
- [Su02] Zhou Su, Jiro Katto, T Nishikawa, M Murakami, T Washizawa and Y Yasuda, An Integrated Scheme for Distributed Segmented Streaming Media over Hierarchical Caches, *International Conference on Information Technology and Applications*, Bathurst Australia, November 2002.
- [Ten00] David Tennenhouse, Proactive Computing, *Communications of the ACM*, vol. 43, No. 5, May 2000.
- [Tua02] N J Tuah M Kumar, S Venkatesh, and S K Das, Performance Evaluation of Speculative Prefetching, *IEEE Transactions on Parallel and Distributed Systems*, Vol. 13, No. 5, June 2002, pp. 471-484.

- [Wan03] Z. Wang, M Kumar, S.K. Das, and H Shen, Investigation of Cache Maintenance Strategies for Multi-cell Environments, 4th International Conference on Mobile Data Management, 21-24 January, 2003, Melbourne, Australia
- [Wei91] M. Weiser, “The computer for the 21st Century,” Sci. American, Sept. 1991.
- [Yue00] Joe Chun-Hung Yuen, Edward Chan, Kam-Yiu Lam and HW Leung, Cache Invalidation Scheme for Mobile Computing Systems with Real-time Data, ACM SIGMOD, Dec. 2000.

Chapter 18

SECURITY IN WIRELESS MOBILE AND SENSOR NETWORKS

Sajal K. Das, Afrand Agah and Kalyan Basu

Center for Research in Wireless Mobility and Networking (CReWMaN)

Department of Computer Science and Engineering

The University of Texas at Arlington

Arlington, TX 76019-0015

Tel:(817)272-7409

{das, agah, basu}@cse.uta.edu

Abstract Security plays a significant role in the current and future wireless mobile networks. The goal of secured communication must ensure that the data we receive is authentic, confidential and have not been altered. Because of the characteristics of wireless medium, limited protection of mobile nodes, nature of connectivity and lack of centralized managing point, wireless mobile networks are too vulnerable and often subject to attacks. In this chapter we discuss security issues and challenges in wireless mobile ad hoc and sensor networks as well as in pervasive computing infrastructures. We also describe security protocols for such environments.

Keywords: Ad hoc networks, authentication, digital signature, intrusion detection, pervasive computing, sensor networks.

1. Introduction

Rapid growth of wireless mobile and sensor networks has changed the perspective of network security altogether. This is due mainly to the characteristics of wireless medium, severe resource constraints as well as unpredictable mobility, all of which contribute to new vulnerabilities that do not exist in a wired network. Therefore, we can no longer use the same mechanisms that were used for Internet security, to protect wireless

mobile and/or sensor networks. First of all, wireless channels are more prone to failures and eavesdropping. Limitations to such resources as CPU, memory, bandwidth and battery power of portable/mobile devices often become a bottleneck in using more sophisticated and computation-intensive key management protocols. Additionally, as mobile nodes are capable of roaming, lack of physical protection results in compromising and capturing these nodes. Clearly, attacks by a compromised node from inside the network is often much more damaging and harder to detect, due to lack of physical protection, than external attacks.

A wireless mobile network is susceptible to a number of compromises both from legitimate users of the system and from intruders. Two general security compromises are host compromise and communication compromise. Host compromises can be categorized as masquerading, unauthorized use of resources, disclosure of information, alteration of information and denial of service. Masquerading is when a user masquerades as another one to gain access to an unauthorized system component. Unauthorized use of resources is when a user accesses system without authorization from real user of the device. Unauthorized reading of stored information is disclosure of information. And unauthorized writing of the information is alteration of information. By denying a resource to an entity, an attacker denies a service. Communication compromises can be categorized as masquerading, interception, fabrication, denial of service, repudiation of actions and unauthorized use of resources. Repudiation of actions is when the sender of a message denies sending it. Interception is when the opponent gains access to the data transmitted over the communication link, so it can obtain the transmitted information or identities and locations of the communication parties.

1.1 Basics of Security Protocols

Any security protocol in a wireless mobile computing environment has the following four goals [5, 11].

Authentication: The portable device (e.g., laptop computer, PDA) used by wireless mobile user and the base station with which the user communicates must be able to mutually authenticate each other. Such authentication enables the base station to prevent unauthorized users from using its services, and enables the user to choose an authorized base station as the access point.

Data privacy: The portable device used by a wireless mobile user and the base station must be able to communicate in a secure way. Data

privacy protects the data from being snooped, replayed or any forgery.

Location privacy: Location privacy is a unique goal of wireless mobile communication. It is desirable or critical not to compromise the location and identity of the mobile user.

Accounting: When a mobile user uses a portable device, it establishes communications with a base station, and the service provider needs to know who to charge for the service used.

Physical layer security mainly aims at preventing every unauthorized access to the devices. There are three ways to authenticate devices: (i) What the user knows; it is very useful for small devices. PIN and password entering fall into this category; (ii) What the user owns; it is accessing the control of some devices in the form of key-cards and electronic tokens. Permanent installed devices could use this type of authentication, like SIM cards in GSM; and (iii) What the user is like; it is also known as biometric authentication and is based on checking the user's physical characteristics. Usually a good authentication scheme involves more than one of the above mechanisms. Like ATM machines require both physical card and a PIN.

The following techniques are the basic building blocks of any security protocol [15].

Symmetric Cryptography: It means that the key for the encryption and decryption of messages is the same. Secret-key cryptography (SKC) scheme is an example that is used for confidentiality of stored data.

Asymmetric Cryptography: It uses algorithms with different keys for encryption and decryption. So both parties that want to transfer data, are able to exchange their encryption keys over an even insecure connection link. The general idea behind public key cryptography is that each participant in a communication network has two keys – one public key and one private key. Agent's public key is freely available, or may be obtained from certifying and trusted authority on demand. The agent's private key is meant to be secret: only agent knows this key and letting others know this key will allow them to assume its identity, as well as decrypt all messages, if intercepted, that were addressed to it. Needless to say, public key and private key are different keys.

Certificate Authority: When two parties exchange their public key over an insecure channel, an attacker can substitute the public keys with fake

ones. One way to verify that the public key really belongs to the sender, is using a certificate authority (CA), who signs the public key with some information about the owner.

The rest of the chapter is organized as follows. Section 2 discusses intrusion detection and highlights some techniques. Section 3 reviews the security aspects in wireless ad hoc networks and different existing solutions. Section 4 talks about security in pervasive computing. Section 5 describes security in sensor networks. Section 6 concludes the chapter.

2. Intrusion Detection

By definition, intrusion is any set of actions that attempts to compromise the integrity, confidentiality or availability of resources [4]. As wireless mobile networks do not have a fixed topology (it changes dynamically due to user and device mobility and also due to fluctuations in wireless links) or fixed infrastructure (as in mobile ad hoc networks), standard intrusion detection techniques that rely on real time traffic analysis, will not be sufficient here. For example, in a wired network we can monitor the network at switches or routers, but in mobile ad hoc networks we do not normally have such traffic to get audit data for the whole network. Moreover, a mobile user may often be disconnected which can be due to the environment and nature of the wireless medium, and may not necessarily represent a model of anomaly. Therefore, a good intrusion detection system for mobile wireless networks must be able to recognize anomaly from normalcy. Zhang, et al. [43, 44] examine the vulnerabilities of wireless networks and developed an intrusion detection architecture for such networks. Traffic flow, routing activities and topology patterns are three general type of information that they consider for their approach. In their system, individual intrusion detection agents are placed at each node of the network. Each agent monitors local activities (user and system). If any evidence of intrusion is found, neighboring agents cooperatively help in a more global intrusion detection.

The main assumption in any intrusion detection scheme is that the program activities are under observation. Based on the system audit data, we can broadly categorize intrusion detection systems into three classes:

- *Network-based intrusion detection:* It runs at the gateway of a network and examines all incoming packets. In a wireless network this can be done at the base station. Intrusion prevention measures such as encryption and authentication can be placed at the base station and any new node will be examined accordingly.

- *Host-based intrusion detection:* It receives the necessary audit data from the operating system and analyzes the generated events. In a wireless network, especially in an ad hoc network, the time duration for which a node acts as a host can be short, and also the topology of the network changes very frequently, so the system must act correctly in the new environment.
- *Router-based intrusion detection:* It is installed on routers to prevent intruder from entering into the network. Intrusion detection system gets audit data from operating system to monitor and analyze the events. Wireless mobile networks do not have traffic concentration points that a system can collect audit data. So all the possible audit data are any data available during the communication within the radio range.

Based on the functional characteristics of detection schemes, intrusion detection systems can be categorized into knowledge based (or *misuse detection*) and behavior based systems (or *anomaly detection*). In misuse detection systems, patterns of well-known attacks or weak spots of the system are used to identify known intrusions. For example, if within two minutes interval there are more than four failed login attempts, the system should raise a flag indicating someone is trying to guess the access code or password. The advantage of this approach is accurate detection of known attacks whereas its disadvantage is the lack of the ability to detect newly invented attacks. For example, if someone has stored his credit card number information in his cell phone, and the cell phone gets lost then anyone who uses this phone would be able to misuse the credit card, without being double checked by the system, until the system learns about this attack.

In anomaly detection systems, if an activity differs from the established normal profile, it is flagged as an anomaly. User's normal behavior profile can contain such information as the amount of network traffic, the daily path that a cell phone user follows everyday, resources used, and so on. The advantage of this approach is that it does not require prior knowledge but its disadvantage is high false positive rate.

With the anticipated growth of wireless Internet technologies, it is expected that the mobile commerce transactions will grow significantly. The security of such a highly dynamic environment is a paramount challenge for the information industry. The major vulnerable components in this infrastructure are password, protocol, content, uncontrolled flow and wireless access links. In the Internet domain, it has been found that most of the common attacks have a target, long-term plan and strategy. Indeed the attackers have their loosely connected global organization and

pursue over longer time to execute the plan successfully. The strategies used to develop the attack plan include [20]: (i) IP spoofing, (ii) SYN flooding, (iii) sequence number attack, (iv) TCP session hijacking, (v) password guessing, (vi) information theft, and (vii) denial of service. Clearly, the wireless Internet infrastructure will also be prone to these attacks, in addition to other attacks that are specific to wireless networks as detailed in the next section. To protect such environment against such malicious activities, it is necessary to design a robust framework based on systematic learning strategies, than solving the symptoms only. Thus securing the wireless Internet is extremely challenging.

3. Security in Wireless Ad Hoc Networks

A wireless ad hoc network is characterized by frequent changes in topology. In other words, the trust relationship between the nodes also changes frequently, which can not be achieved with static security configurations. Moreover, the security configuration must be scalable to accommodate hundreds of nodes in the ad hoc network. Routing protocols must not only be able to cope with dynamically changing topology but also malicious attacks. Since hosts may move in an arbitrary manner in such networks, the routes are subject to frequent disconnections and re-connections.

Traditionally, to secure a wireless network, we consider a wide range of issues: (i) identification of a mobile user, (ii) anonymity of a mobile user, (iii) authentication of a base station, (iv) security of the flow information, (v) prevention of attack, (vi) computational complexity of having authenticity and security at mobile terminal, and (vii) cost of establishing a session key between user and base station.

Getting access to wireless (radio) links is relatively easy, implying the security of wireless communications can be easily compromised, by interception or jamming of radio signals. In particular, if transmissions occur in a large area or when the users are allowed to cross security domains, like permission to use one service in one environment but prohibiting the use of another service in the same area. By using wireless networks we are able to use more flexible communication models than traditional wire-line networks. Mobility and volatility are transparent from the application so each node can communicate with any other node, as it would be in a fixed wired network. In wired networks an attacker must have physical access to the network or passes several fire-walls and gateways, but attacks against a wireless ad hoc network can come from many directions. An attacker may try to steal information, contaminates messages or impersonate a node. Nodes that are roaming in a

hostile environment and have poor physical protection have a probability of being compromised, so attacks from inside the network must be considered as important as attacks from outside the network.

If we are to list some of the threats and vulnerabilities of wireless ad hoc networks, the following attacks are the most common ones: (i) all vulnerabilities that exist in a wired network can also be present in wireless networks, (ii) information that is not encrypted and is transmitted between two wireless devices can be intercepted, (iii) malicious user can gain unauthorized access and bypass any firewall protection, (iv) handheld devices with sensitive information can be easily stolen, and (v) data can be corrupted or jammed.

3.1 Digital Signature Based Routing

In most proposed routing protocols for mobile ad hoc networks, routers exchange information to establish routes between different nodes. This information can be very inviting to attackers. Introducing excessive traffic load into the network or advertising incorrect routing information to other nodes are different kinds of attacks to routing protocols.

Dahill, et al. [14] and Sanzgiri, et al. [34] present a secure routing protocol (ARAN) which requires the use of a trusted certificate authority (CA); before a node enters in an ad hoc network, it must get a signed certificate from the CA. The route discovery process is initiated by the source node by flooding a digitally signed Route Discovery packet to its neighbors. When a neighbor receives this message, it sets up a reverse path back to the source node and verifies the signature. ARAN provides end to end and hop by hop authentication of route discovery and reply messages. The certificate includes IP address of node, public key and a time stamp showing when the certificate is created. All nodes are supposed to maintain fresh certificates with the trusted server. This protocol first verifies that the destination is reached, then verifies that a route discovery process has found the shortest path. A source initiates the route discovery process to reach to the destination, by broadcasting a route discovery packet to all its neighbors. Eventually the packet reaches the destination, and the destination node unicast a reply to the source. There is no guarantee that this first received packet at the destination has traversed the shortest path, so the source also broadcasts shortest path confirmation message to its neighbors.

To achieve a highly secure key management service, Zhou and Haas [45] have proposed threshold cryptography, to distribute trust among a set of servers. In this approach a set of arbitrary nodes are chosen as servers. All nodes know the public key but the private key is divided into

n shares between n servers. As shown in Figure 18.1 , each server has a public/private key pair K/k . Private key k is divided into n shares, one for each server. Each server also has a public/private key pair K_i/k_i and knows the public keys of all nodes. A $(n, t + 1)$ -threshold cryptography scheme allows n parties share the ability to perform a cryptographic operation so that any $t + 1$ parties can perform this operation jointly, but it is infeasible to do by at most t parties. The drawback of this mechanism is that in mobile ad hoc networks we rather have the ability to work without any external management or configuration, but the proposed mechanism relies on a subset of nodes playing specific roles at a given point of time, which is undesirable.

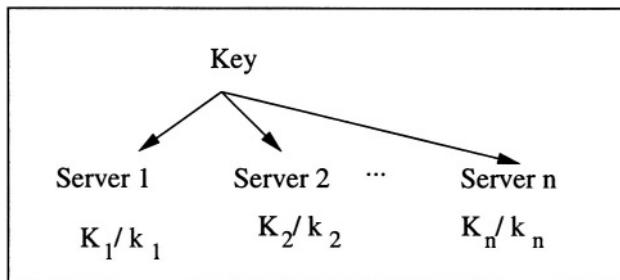


Figure 18.1. Key Distribution

Hubaux, et al. [12, 21] propose a self-organizing public-key management system for mobile ad hoc networks. Certificates are stored and distributed by the users. Each user maintains a local certificate repository, which contains a limited number of certificates. When two users want to verify the public key of each other, they merge their repositories and find a certificate chain, which makes the verification possible. This functionality of this approach depends on the construction of local certificate repositories and certificate graph. A certificate graph is a graph whose edges represent public key certificates issued by the users and whose vertices represent public keys of the users. Each user has several directed and mutually disjoint paths of certificates in its local repository. Each path starts at the node itself and certificates are added to this path such that a new certificate is chosen among those connected to the last node on the path, and the new certificate leads to the node with the highest number of certificates connected to it. When verifying a certificate chain, a node must trust the issuer of the certificate in the chain for checking that the public key in the certificate really be-

longs to the node named in the certificate. A drawback of this approach is that before performing authentication, each node must build a local repository which is relatively expensive.

Hu, et al. [22] present a secure routing protocol, based on the Destination Sequenced Distance Vector protocol (DSDV). Each node updates its table using the updates it hears, so its route for each destination uses as a next hop the neighbor that advertised the smallest metric for that destination. In this protocol, called secure efficient ad hoc distance vector (SEAD), nodes periodically exchange routing information with each other, so each node always knows a current route to all possible destinations. The protocol uses efficient one way hash chains rather than relaying by using expensive asymmetric cryptography operations. It deals with attacks that modify routing information broadcasted during the update phase of the DSDV protocol. The SEAD protocol assumes a mechanism for a node to distribute an authentic element of the hash chain that can be used to authenticate all the other elements of the chain. This protocol authenticates the sequence number of routing table update messages using hash chains elements. Each node uses an element from its hash chain in each routing update. Based on this element, the one-way hash chain provides authentication in other routing updates for that node.

In [23], Hu, et al. present an on-demand secure ad hoc routing protocol (ARIADNE), based on the Dynamic Source Routing (DSR) protocol, which relies on symmetric cryptography. They present a mechanism that enables the target to verify the authenticity of the ROUTE REQUEST. A per-hop hashing technique is used to verify that no node is missing from the node list in the REQUEST. This protocol guarantees that the target node of a route discovery process authenticates the initiator. To convince the target of the authenticity, the initiator includes a unique data (such as time-stamp) in each packet. The initiator can authenticate each intermediate node to the destination, using PREP message, and no intermediate node is able to remove a node from the node list in the PREQ or PREP messages. When the target wants to authenticate nodes, it will return a REPLY message on a path that only contains legitimate nodes. One drawback of ARIADNE is the high cost of key setup.

3.2 Enforcing Cooperation

Buttyan, et al. [8, 9] present packet purse model. In this model each packet is loaded with nuglets by the source, as depicted in Figure 18.2 , and each forwarding host takes out nuglets for its forwarding

service. When a node wants to send a packet as originator, the number of forwarding nodes that are needed to reach the destination is estimated. When a node forwards a packet, its nuglet counter is increased by one. If a node wants to send its own packets, then it must forward packets for benefit of other nodes. So this approach discourages users from flooding the network but the downside is that the source needs to know exactly how many nuglets it has to include in the packet it sends. Another problem with this approach is that intermediate nodes can take out more nuglets than they are supposed to. The solution is based on using nuglet or virtual currency. Another type of nuglet that nodes can use is packet trade. In the packet trade model, each packet is traded for nuglets by the intermediate nodes as follows: each intermediate node buys the packet from the previous node on the path. So the destination has to pay for the packet. Therefore, the source does not need to know how many nuglets are to be loaded in the packet; however malicious flooding can not be prevented either.

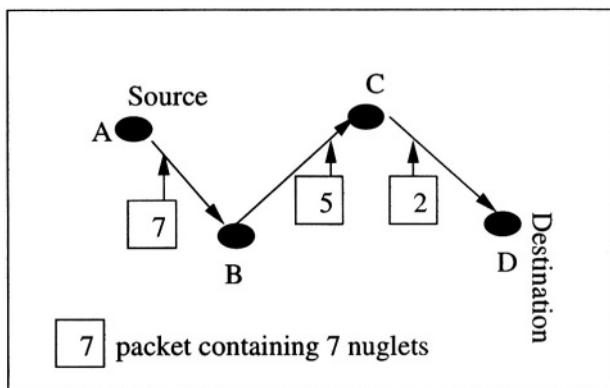


Figure 18.2. Nuglet model

Buchegger and Boudec [6] propose an approach to find selfish and malicious nodes. (A selfish behavior is one that maximizes the utility of a node.) The protocols, called, cooperation of nodes and fairness in dynamic ad hoc networks (CONFIDANT), observes nodes and detects attacks. In this protocol, nodes monitor their neighbors and if there is reason to believe that a node misbehaves, they take action in terms of their own routing and forwarding messages. This protocol encourages cooperation by allowing selfish nodes an incentive to cooperate, which leads to an outcome that is beneficial for the network.

The user cooperation in ad hoc networks has also been studied in [36], which proposes an acceptance algorithm that each node uses, to decide whether to accept or reject a relay request. The system is proved to converge to a Nash equilibrium point in game theoretic sense. Srinivasan, et al. assume that each user has sufficient information about the system, like the number of users in each energy class, and hence users exchange their view of the system. However, they do not consider malicious users. A framework to study the existence of cooperation in packet forwarding in a wireless network is proposed in [18], in which a model is defined and the conditions under which cooperative game strategies can form an equilibrium are identified. This approach does not require each node to keep track of the behavior of other nodes, but it is assumed that all routes are static.

3.3 Denial of Service Attacks

The denial of service (DoS) is an attack that prevents the use of networks by even legitimate users. Recently, DoS attacks have become serious security threats. This is due to the fact that, they can result in massive service disruption. Although it is possible to attack software vulnerabilities, yet DoS attacks usually consume finite resources such as bandwidth, CPU processing power or memory. As nodes use wireless communication, one possible kind of attack on the physical layer is jamming. An intruder may corrupt the network with the radio frequencies that nodes in the network are using. Constant jamming prevents a node to exchange data or even report the attack. If the intruder can jam the entire network, then nodes can go to sleep mode and wake up periodically and check if the jam has ended. Cellular phone networks use code spreading as a defense to DoS attacks [40].

Defense against DoS attacks is hard as the IP address of source is spoofed, which makes it difficult to determine the true origin. There are two kinds of DoS attacks, which use packet floods to consume resources [41].

- *Network resource attack:* when an attacker sends useless packets to a victim server with the intention of wasting bandwidth connecting to the server.
- *Server resource attack:* when an attacker sends useless packets with intention of corrupting the ability of server to process the load, or even taking advantage of ambiguities in protocols, like reservation of resources for half-open connections like SYN flooding.

Yaar, et al. [41] propose a packet marking approach where a path fingerprint is embedded in each packet, thus enabling a victim to identify packets traversing the same paths. They observed that reconstruction of the exact path to the attacker is not necessary. As path information is in each packet, each node can filter packets based on the information carried.

4. Security in Pervasive Computing

A pervasive computing environment includes numerous secured, often transparent computing devices, which are mobile or embedded in the environment and are connected to an increasingly ubiquitous network infrastructure. A pervasive computing environment is usually more knowledgeable about the users' behavior. And as time passes, the environment becomes more proactive with each individual user [27]. So the user must be able to trust the environment and also the environment must be confident of user's identity. Security is an important concern in the success of a pervasive computing system which must thus be aware of its user's state and surroundings, and must be able to modify its behavior based on this information. A key challenge is to obtain such information. In some cases the information could be part of the user's personal computing space.

Satyanarayanan [35] emphasizes the following research thrusts in the area of pervasive computing.

Effective use of Smart Spaces: The space may be an enclosed area, like a meeting room or a well-defined open area, like a campus. By embedding computing infrastructure, a smart space brings two disjoint worlds together, for examples, adjustment of lighting levels in a room based on the occupant's profile.

Invisibility: This deals with complete disappearance of pervasive computing technology from a user's consciousness. A good tool must be invisible and make us focus on the task and not on the tool. Eyeglass is an example of a good tool through which we look at the world, not the eyeglasses. According to Weiser [39] "*The most profound technologies are those that disappear. They weave themselves into the fabric of everyday life until they are indistinguishable from it.*"

Localized scalability: As smart spaces grow, the interactions between a user's personal computing space and his/her surrounding will grow too.

So it has bandwidth and energy distraction for a wireless mobile user.

Masking uneven conditioning: The rate of excerpt of pervasive computing into infrastructure will depend on many non-technical factors like economics and business models. And so when users get used to some level of smartness in one environment, smartness-less in another environment would be very visible and unwanted.

Pervasive computing covers a wide spectrum of applications – from switching on the lights in a conference room to buying and selling stocks, booking airline tickets, providing health-case and telemedicine services, emergency rescues, crisis management, and many more complex ones. This has the potential to create a truly “smart environment”.

4.1 Distributed Trust

Pervasive computing involves the interaction, corporation and coordination of several transparent computing devices. One way to satisfy the requirements of pervasive computing is adding distributed trust to the security infrastructure [37]. This approach needs to (i) articulate policies for user authentication and access control; (ii) assign security credentials to individuals; (iii) allow entities to modify access rights of other entities; and (iv) provide access control. Role based control is probably one of the best known methods for access control, where entities are assigned roles, and there are rights associated with each role. This is difficult for systems where it is not possible to assign roles to all users, and foreign users are quite common. Security systems should not only authenticate users, but also allow them to delegate their rights. Agents are authorized to access a certain service if they have the required credentials.

Bharghavan [5] uses additional ontologies that include not just role hierarchies but any properties expressed in a semantic language. For example if an agent in a meeting room is using the projector, it is possible that the agent is a presenter, and must be allowed to use the computer too. So they assign roles dynamically and without making new roles. They extent Agent Management System (AMS) and the Directory Facilitatory (DF) to manage security of the platform. The AMS, DF and agents will follow some security policies to give access rights. They use PKI (public key infrastructure) handshaking protocol between AMS and agents to verify both parties. Once an agent signs a request it will be accountable for it. When an unknown agent attempts to register with a platform, the platform checks the agent’s credential and makes a decision. In multi-agent systems, as they are decentralized, it is impossible

to have a central database of access rights. Therefore, the policy is checked at two levels: (i) controlling access to the AMS and DF, and (ii) specifying who can access the service. Security can be classified into two levels – platforms and agents. In platform security, the AMS and DF decide to allow an agent to register, search or use other functions. An agent can send security information to the AMS and specifies its category as private, secure or open. When an agent wants to register with the AMS, it signs its request and sends with it the digital identity certificate to the AMS. If it is valid, then the AMS decides about the access rights. If the agent does not have the right to register with the AMS, the latter starts a handshaking protocol.

4.2 Incorporating Vigil System

Based on role based access control with trust management, the Vigil system [25] has five components: (i) service broker, (ii) client, (iii) certificate controller, (iv) role assignment manager, and (v) security agent. The vigil system is divided into *Smart Spaces*, each controlled by a broker. The broker finds matching service for the user. All clients must register with the broker. Each client is only concerned with the trust relationship with its broker. The service brokers establish trust relationships with each other. Trusts between clients are transitive through brokers. When a client registers with a broker, it transmits its digital certificate, a list of roles which can access it, and a flag indicating if the broker should publish that client's presence to other clients. After receiving the registration, the broker verifies the client's certificate in order to send it the digital certificate. The communication manager is a bridge between the broker and the entities in the Smart Space, hence it translates protocols. The certificate controller generates digital certificates.

To get a certificate, an entity sends a request to the certificate controller which signs it and then sends back with its own signed certificate. The role assignment manager maintains access control list for entities. An entity can have more than one role. The security agent manages trust in the Space. All security agents will enforce the global policy and a local policy, which is related to the Space. At the beginning, it reads policies and stores them in a database. All permissions and prohibitions are defined there. Access rights can be delegated, they must be from an authorized entity to another entity. When a user needs to access a service, which does not have the right to do that, it asks another user who has that right. If the asked entity has the permission to delegate the access, it sends a message signed by its own private key with its

certificate to the security agent. The security agent verifies that the delegator has the right to delegate, then sets a short period of validity for that permission. After expiring of this period, the delegation must be reprocessed.

Consider a telemedicine scenario as an application for pervasive computing. Suppose Jack has heart attack history, and his doctor wants to have up-to-the minute information about his heart. Jack will use a heart monitor unit, which can be connected to his cell phone. Information will be transferred to his doctor's computer, and accordingly Jack can get diagnostic notification from his doctor. In this case, if someone alternates Jack's information or impersonate the doctor it could have fatal results. Jack and the doctor should not be able to send some information and then deny sending. Jack's information is his privacy; his insurance company has it, but his boss should not be able to access it, so any unauthorized access is forbidden. Having this scenario in mind, let us see how to incorporate security mechanisms in pervasive computing systems.

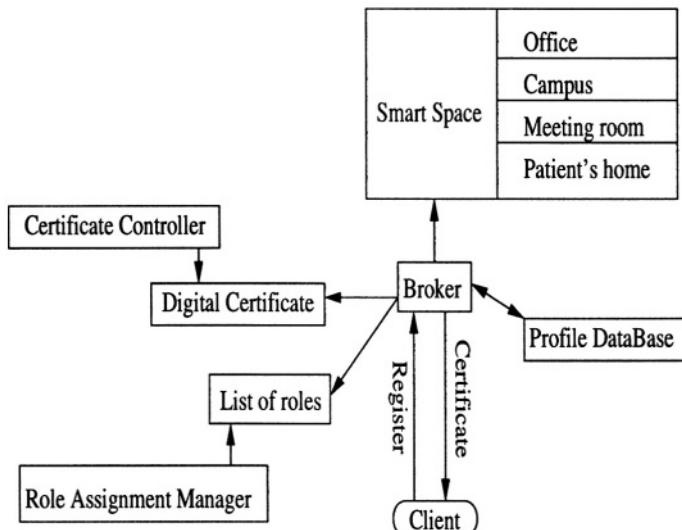


Figure 18.3. Incorporating VIGIL system for a Smart Space scenario

As Figure 18.3 illustrates, we can have different *Smart Spaces*. Each client must register through a broker who gets the digital certificates from the certificate controller and has access to the list of roles provided by the role assignment manager. Then the broker issues certificate to the

client. The broker also has access to a profile database. Each different client has a different profile based on his/her previous interaction with the environment. As discussed in [24, 37], we can use the distributed trust concept so entities can modify access rights of other entities by delegating their access rights. It is also possible to have a chain of delegation and then revoke them accordingly. It is worth mentioning that when one entity gives one right to another entity, the entity is able to access services without system creating a new identity for it.

5. Security in Wireless Sensor Networks

A wireless sensor network consists of different kinds of sensors that monitor the environment. Due to low cost and flexibility, sensor networks are being rapidly deployed for a variety of applications such as ecology and environment monitoring, military, emergency sites, remote sensing, smart homes, health, and commercial areas [3]. However, sensors have very limited resources (CPU, memory, bandwidth, batter power). Thus, designing resource-efficient protocols is a major challenge in sensor network design. Other issues include security, fault-tolerance, survivability, scalability.

As sensor networks occur in the real world, there is a need to protect them from various types of attacks. This is certainly challenging in an environment where the network is designed to be flexible. It is crucial that the security of sensor networks be monitored and diagnosed to ensure their proper behavior. The major challenges in tackling sensor networks security include power conservation for mobile sensors, cooperation among heterogeneous sensors, flexibility in the security level to match the application needs and conserve critical resources, scalability, self organizing and self learning of sensors, trust and security decisions for the application, yet keeping the mobility and volatility transparent and protecting the network from external and internal intrusions. In a nutshell, we must consider three important factors – energy, computation and communication.

Due to resource scarcity of sensors, securing sensor networks is quite different and challenging from traditional schemes that generally involve management and safe keeping of a small number of private and public keys [31]. Disclosing a key with each packet requires too much energy [32]. Storing one-way chain of secret keys along a message route requires considerable memory and computation in the nodes on that route [33]. The key management using a trusted third party again requires an engineered solution that makes it unsuitable for sensor network applications [31]. Although the asymmetric key cryptography does not require

a trusted server, key revocation becomes a bottleneck [31] as it involves an authority maintaining a list of revoked keys on a server or requesting the public key directly from the owner.

As mentioned earlier, there are two type of attackers, namely clever or curious outsiders and knowledgeable insiders. An attacker can take over any node within a network. Subverting a single node means that all nodes with the communication radius of that node can be denied receiving any information. Therefore, the goal is to minimize the impact of a subverted node on the rest of the network. The single node should not grant the attacker the ability to subvert the entire network. In a network with a single gateway (i.e., base station), the attacker needs to take over just that one node to render the network inactive. In order to be able to trust sensor data, the source must be authenticated so the malicious sensors can not inject false data. For checking the integrity of data, we must be able to detect data modification. Additionally, data must be confidential so no one else can read it. Thus the threats that a sensor network may encounter are eavesdropping, message injection, message replaying, message modification and denial of service.

The security primitives are message confidentiality, authentication, message integrity, as message freshness are summarized below [31].

Data Confidentiality: A sensor network should not leak sensor readings to other networks. Usually nodes communicate sensitive data. In order to protect sensitive data, they are encrypted with a secret key that only intended receivers have. Moreover, secure channels are set up between the nodes and base stations.

Data Authentication: Since an adversary can easily inject messages, there must be a way that assures receiver about the origin of the data. By authenticating the data we allow the receiver to verify that the data was sent from the claimed sender. In case of two-party communication, the sender and receiver can share a secret key. But this method can not be used in broadcasting as anyone can impersonate the sender and forge messages to other receivers.

Data Integrity: It ensures that the data have not been altered by any adversary while in transit.

Data Freshness: It implies that the data is recent and no adversary replayed old messages.

In most sensor networks, some nodes have special responsibilities. They can act as a coordinator, monitoring access point or cryptographic key manager [29]. So they are more attractive for intruders, as they can provide access to network services. A passive intruder can observe the network traffic and learns the location of these critical resources and act later; this is known as *homing attack*. If all nodes use cryptographic keys, the network encrypts the header of each packet at each hop, and prevents intruder from easily learning about the network. Another kind of attack is *desynchronization*. The intruder forges messages to one or both end points. These messages carry wrong sequence numbers and cause the end points to ask for retransmission of lost packets. Hence a clever intruder can prevent the two end points from exchanging any useful data.

5.1 Sensor Security Schemes

Researchers in UC Berkeley have designed two security building blocks, called SNEP (Secure Network Encryption Protocol) and μ TESLA (“micro” version of Timed, Efficient, Streaming, Loss-tolerant Authentication Protocol) [33].

5.1.1 SNEP Protocol. The SNEP protocol provides data confidentiality, two-party data authentication, data integrity and freshness of data. Data confidentiality is a security primitive which is used in every security protocol. It has low communication overhead since it only adds 8 bytes per message. Also, like many other protocols, it uses a counter. It has semantic security, which ensures an eavesdropper has no information about the plain-text, even if it sees multiple encryptions of the same plain-text [32]. The SNEP protocol uses a cryptographic mechanism that achieved semantic security with no additional transmission overhead. It relies on the value of the counter which is shared between the sender and receiver, and this value is incremented after each block of cipher message. Combination of this and the Message Authentication Code (MAC) forms the SNEP protocol. As the counter is kept at each end point and is not part of the message, it has low communication overhead. The counter value in the MAC also prevents replaying of old messages. In the absence of a counter value in MAC, any adversary may be able to replay old messages [32].

5.1.2 μ TESLA Protocol. The μ TESLA provides authentication for data broadcasting. It requires the base station and nodes be loosely time synchronized and each node knows an upper bound on the maximum synchronization error. The base station simply computes a

MAC on the packet with a secret key. When a node receives a packet, it verifies that the corresponding MAC key was not yet disclosed by the base station. The receiving node ensures that the MAC key is known by the base station so no adversary could have altered the packet. The node then stores this packet in a buffer. At the time of key disclosure, the base station broadcasts the verification key to all the receivers. When a node receives a disclosed key, it can verify the correctness of it. If it is correct, the node uses it to authenticate the buffered packet. The μ TESLA protocol provides broadcast mechanism, uses asymmetric mechanism to prevent forgery and does not use asymmetric digital signature as it is computationally expensive and needs more storage.

The sender and receiver share the initial TESLA key as follows. The receiver can send a message to base station, asking for the current TESLA key over a secure channel. Since the reply from the base station is encrypted and authenticated, the receiver can bootstrap into the TESLA group. Also by fine-tuning the disclosure interval, limited buffering is needed. In this protocol the sender first generates a sequence of secret keys. The sender chooses the last key randomly, and generates the remaining values by applying a one-way function on it. As time is divided into time intervals, the sender associates each key of the one-way key chain with one time interval. The sender uses the key of current interval to compute MAC. Then it reveals the key after some intervals. The important property of one-way key is that once the receiver has an authenticated key of the chain, subsequent keys of the chain are self-authenticating. Due to loose time synchronization and authenticated key chain, the freshness and point-to-point authentication can be provided. The MAC uses the secret key shared by nodes and the base station, thus authenticating the data. When the receiver receives the packet, it needs to make sure that the sender did not disclose the key of incoming packet. This asymmetric delayed key disclosure requires loosely synchronized clocks between the sender and the receiver.

5.1.3 Key Management. The limited resources make it undesirable for sensor nodes to use public-key algorithms, such as Diffie-Hellman key agreement [16]. A sensor node may need tens of seconds or even minutes to perform these operations. Usually a large number of sensor nodes are deployed in hostile environments, necessitating them to be low-cost. At the same time it is harder to make them tamper-resistant [17]. An adversary can take control of a sensor node and compromise the cryptographic keys. It is very likely that a sensor node does not know which other nodes will be coming within its communication range. In other words, there is no *a priori* knowledge of neighboring sen-

sor nodes. Due to limited memory, battery power and other resources, a sensor node can not establish a unique key with every other node in the network. Moreover, a sensor network is vulnerable not only to internal attacks from untrustworthy nodes, but also to external attacks from sources like mobile codes that are used for activating remote software upgrade, configuration management, dynamic resource monitoring, etc. In such an environment any node could claim that it is one hop away from the destination, causing all routes to the destination to pass through. Also a malicious node could corrupt any packet and cause data to be misrouted.

As sensor networks can be deployed in hostile areas, where each node is subject to be captured by adversary, they require protection over their communications. A key-management scheme is proposed in [13], which is designed to satisfy security requirements for the network. This scheme consists of different phases. They generate a large pool of keys, randomly draw a key, load the key in the memory of each sensor node, and save the key identifier on a trusted controller. Each node must discover its neighbors. If a node is compromised, the entire key must be revoked. All such communications need energy, which is very limited to sensor nodes. Meanwhile, the limited computing power is a bottleneck for using more sophisticated protocols.

5.2 Enforcing Cooperation and Reputation

This section describes a game theoretic framework due to Agah, Das and Basu [1], to enforce cooperation and reputation in wireless sensor networks.

Game theory offers a formal tool to analyze interaction among a group of rational players who behave strategically. A game is the interactive situation, specified by the set of players (sensor nodes), the possible actions of each node, and the set of all possible payoffs. What one individual player does directly affects at least one other player in the group. Games in which the actions of the players are directed to maximize the profit without subsequent subdivision of the profit among the players are called *cooperative games*. In such games, the outcome is a result of an agreement among players. These games are compared with respect to the preferability of payoffs [38]. In other words, in a cooperative game, different players form alliance with each other in a way to influence the outcome of the game in his/her favor. In *noncooperative games*, however, no outside authority assures that players stick to the same pre-determined rules, and binding agreements are not feasible. It is known that in noncooperative games, there exist sets of optimal strategies, the

so called Nash equilibrium [38], used by the players in a game such that no player can benefit by unilaterally changing his or her strategy if the strategies of the other players remain unchanged.

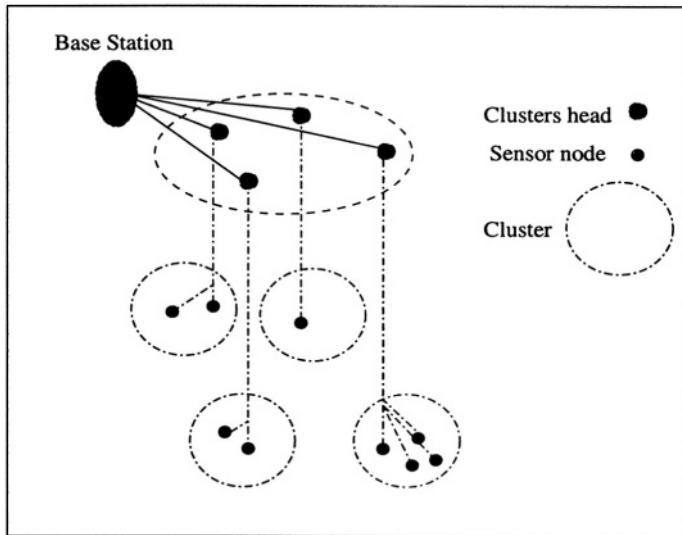


Figure 18.4. Structure of a Sensor Network

5.2.1 Game and Payoff Function. In [1], a cooperative game is defined between the sensor nodes. A set of strategies for each node is chosen and the payoff function for each node is also defined. Any subset of players (i.e., sensor nodes) will be called a cluster (coalition), as shown in Figure 18.4. The goal of the game is to find the largest payoff that sensor nodes within a cluster can collectively achieve, assuming that all nodes in one cluster use the same key. All sensors play the game in such a way as to have a more populated cluster. But, on the other hand, the total possible number of sensor nodes in each cluster can not be more than what the cluster can handle (this is due to resource constraints). Thus an optimal payoff function is defined based on the security related objectives in the sensor network. This function consists of three factors: (i) cooperation, (ii) reputation and (iii) quality of security.

The payoff between two sensor nodes depends on their distance and each node's transmitter signal strength. The more the transmitter signal strength, the more likely is its cooperation with the close neighbors. In order to determine how much each individual node is useful for the

whole network (a measure of reputation), the payoff between should also represent how many packets each node receives and forwards at each time slot. Finally, the payoff between two sensor nodes should also represent the trustworthiness of the traffic. In each cluster, one node with the highest cooperation, best reputation and highest quality of security is chosen as a clusterhead.

5.2.2 The Strategy Set. In each time slot, a sensor node uses its strategy based on the information obtained in preceding time slots. Consider a packet is coming from a source to a destination node, passing through an intermediate node. This intermediate node decides to forward or not to forward the packet to the destination node, based on three following criteria:

- Reputation: Have these two sensor nodes made enough reputation to trust each other and to cooperate with each other, in regards to forwarding packets on behalf of each other?
- Distance: The closer the two nodes together, the more they trust each other. If they are too far away from each other, it is less likely that they will trust each other.
- Traffic: Have these two sensor nodes a good history of joint operation, and can they trust each other?

If there is enough reputation level, closeness, and good history of joint operation, then the strategy for sensor nodes is to cooperate with each other (where cooperation means receiving or forwarding incoming packets); otherwise the strategy is to defect. For details, refer to [1].

6. Conclusions

Migration of information from mainframe to laptops introduces more stringent security requirements into our daily lives. Wireless devices can not be treated as gadgets and security must be considered in the design of transport layer, network layer as well as in data link layer. In this chapter we discussed various security related issues, including intrusion detection and denial of service, and surveyed various security protocols in wireless mobile networks such as ad hoc and sensor networks as well as pervasive computing environments. Clearly, due to the nature of each of these networks, a well defined protocol in one infrastructure can not be very useful in another one. We believe security should be taken into account at the early stages of the network design. Furthermore, the security solutions proposed for wireless mobile and sensor networks must consider

the limitations on battery life, computing, wireless communication, and storage resources of network components.

Acknowledgment

This work is supported by NSF ITR grant under Award Number IIS-0326505.

References

- [1] A. Agah, S. K. Das and K. Basu, "A game theory based approach for security in sensor networks", submitted to IPCCC 2004.
- [2] Agah, S. K. Das and K. Basu, "Security in Sensor Networks", submitted to WCNC 2004.
- [3] I. F. Akyildiz, W. Su, Y. Sankarasubramanian, and E. Cayirci, "Wireless Sensor Networks: A Survey", *Computer Networks*, vol. 38, pp:393-422, 2002.
- [4] B. Balajinath and S. V. Raghavan, "Intrusion detection through learning behavior model", *Computer Communications* vol.24, pp:1202-1212, 2001.
- [5] V. Bharghavan, "Secure Wireless LANs". *Proceedings of the ACM Conference on Computers and Communications Security*, pp:10-17, Virginia, 1994.
- [6] S. Buchegger and J. Le Boudec, "Performance analysis of the CONFIDANT protocol", In *proceedings of the 3rd ACM International Symposium on Mobile Ad Hoc Networking and Computing (Mobi-Hoc)*, pp:226-236, June 2002.
- [7] L. Buttyan, J. P. Hubaux and S. Capkun, "A Formal Analysis of Syversons Rational Exchange Protocol" , *CSFW*, pp:181-193, June 2002.
- [8] L. Buttyan and J. P. Hubaux, "Stimulating Cooperation in Self-Organized Mobile Ad Hoc Networks", *ACM/Kluwer MONET*, vol.8, no.5, October 2003.

- [9] L. Buttyan and J. P. Hubaux, "Nuglets: a virtual currency to stimulate cooperation in self-organized ad hoc networks", Technical Report, 2001.
- [10] M. Bykova, S. Ostermann and B. Tjaden, "Detecting Network Intrusions via a Statistical Analysis of Network Packet Characteristics", *IEEE Southeastern Symposium on System Theory*, pp:309-314, 2001.
- [11] R. H. Campbell, T. Qian, W. Liao and Z. Liu, "Active Capability: A Unified Security Model for Supporting Mobile, Dynamic and Application Specific Delegation", *Security White Paper: System Software Research Group*, Department of Computer Science, University of Illinois at Urbana-Champaign, 1996.
- [12] S. Capkun, L. Buttyan, and J. P. Hubaux, "Self-Organized Public Key Management for Mobile Ad hoc Networks", *IEEE Transactions on Mobile Computing*, vol. 2, no. 1, June 2003.
- [13] H. Chan, A. Perrig and D. Song, "Random Key predistribution Schemes for Sensor Networks", In *Proceedings of the 2003 IEEE Symposium on Security and Privacy*, pp:197-213, 2003.
- [14] B. Dahill, B. N. Levine, E. Royer and C. Shields, " ARAN: A secure Routing Protocol for Ad Hoc Networks", UMass Tech Report, 2002.
- [15] Y. Desmedt, "Threshold Cryptography", *European Transactions on Telecommunications*, vol.5, no.4, pp:449-457, July 1994.
- [16] W. Diffie and M. E. Hellman, "New directions in cryptography", *IEEE Trans. Inform. Theory*, pp:644-654, November 1976.
- [17] L. Eschenauer and V.D. Gligor, "A Key-management Scheme for distributed Sensor Networks", In *Proceedings of ACM CCS*, pp:41-47, Washington, DC, November 2002.
- [18] M. Felegyhazi, L. Buttyan and J.P. Hubaux, "Equilibrium Analysis of Packet Forwarding Strategies in Wireless Ad Hoc Networks-the Static Case", *Proceedings of Personal Wireless Communications (PWC '03)*, Venice, Italy, September 2003.
- [19] B. Gelbord, "Graphical Techniques in Intrusion Detection Systems", *IEEE 15th International Conference of Information Networking*, pp:253-258, 2001.

- [20] B. Harris and R. Hunt, "TCP/IP security threats and attack methods", *Computer Communications*, vol.22, no.10, pp:885-895, June 1999.
- [21] J. P.Hubaux, L. Buttyan and S. Capkun, "The Quest for Security in Mobile Ad Hoc Networks", *Proceedings of the ACM Symposium on Mobile Ad Hoc Networking and Computing (MobiHOC)*, pp:146-155, October 2001.
- [22] Y. C. Hu, D. B. Johnson and A. Perrig, "SEAD: Secure Efficient Distance Vector Routing for Mobile Wireless Ad Hoc Networks", in *Proceedings of the 4th IEEE Workshop on Mobile Computing Systems and Applications (WMCSA)*, pp:3-13, June 2002.
- [23] Y. C. Hu, A. Perrig and D. B. Johnson, "Ariadne: A Secure On-demand Routing Protocol for Ad hoc Networks", *MobiCom*, pp:12-23, 2002.
- [24] L. Kagal, J. Undercoffer, F. Perich, A. Joshi and T. Finin, "A Security Architecture Based on Trust Management for Pervasive computing Systems", *Conference - Grace Hopper Celebration of Women in Computing*, 2002.
- [25] L. Kagal, J. Undercoffer, A. Joshi and T. Finin, "Vigil: Enforcing Security in Ubiquitous Environments", <http://www.csee.umbc.edu/lkagal/papers/vigil.pdf>.
- [26] M. Kodialam and T.V. Lakshman, "Detecting Network Intrusions via Sampling: A Game Theoretic Approach", *INFOCOM*, pp:1880-1889, 2003.
- [27] M. Kumar, B. Shirazi, S. K. Das, B. Sung, D. Levine, and M. Singhal, "PICO: A Middleware Framework for Pervasive Computing", *IEEE Pervasive Computing*, vol. 2, no. 3, pp:10-17, July-Sept 2003.
- [28] L. Lankewicz, M. Benard, "Real-time Anomaly Detection Using a Non-Parametric Pattern Recognition Approach", *Proceedings 7th Annual Computer Security Application Conference*, pp:80-89, 1991.
- [29] S.J. Lee, W. Su and M. Gerla, "Mobility Prediction in Wireless Networks", In *Proceedings of IEEE ICCCN*, vol. 1, pp:22-25, October 2000.
- [30] P. Papadimitratos and Z.J. Haas, "Secure Routing for Mobile Ad hoc Networks", *CNDS*, pp:27-31, January 2002.

- [31] A. Perrig, R. Canetti, J. D. Tygar and D. Song, "Efficient Authentication and Signing for Multicast Streams over lossy Channels", In *IEEE Symposium on Security and Privacy*, pp:56-73, May 2000.
- [32] A. Perrig, R. Szewczyk, V. Wen, D. Culler and J. D. Tygar, "SPINS: Security Protocols for Sensor Networks", *MobiCom*, pp:189-199, July 2001.
- [33] A. Perrig and J. D. Tygar, "Secure Broadcast Communication in Wired and Wireless Networks", *Kluwer Academic Publisher*, ISBN 0792376501, 2003.
- [34] B. Sanzgiri, B. Dahill, B. N. Levine, C. Shields and E. M. Belding-Royer, "A secure routing protocol for ad hoc networks", In *Proceedings of the 10th IEEE International Conference on Network Protocols (ICNP)*, pp:12-15, November 2002.
- [35] M. Satyanarayanan, "Pervasive Computing: Vision and Challenges", *IEEE Personal Communications*, pp:10-17, August 2001.
- [36] V. Srinivasan, P. Nuggehalli, C. F. Chiasserini and R. R. Rao, "Cooperation in Wireless Ad Hoc Networks", *INFOCOM*, vol. 2, pp:808-817, 2003.
- [37] J. Undercoffer, F. Peirch, A. Cedilink, L. Kagal and A. Joshi, "A Secure Infrastructure for Service Discovery and Access in Pervasive Computing", *ACM Journal of Mobile Networks and Applications*, Special issue on security in Mobile Computing Environments, vol. 8, no. 2, pp:113-125, April 2003.
- [38] N. N. Vorobev, "Game Theory lectures for economists and systems scientists", *Springer- Verlag*, ISBN 0387902384, 1977.
- [39] M. Weiser, "The Computer for the 21st Century", *Scientific American*, vol. 265, no. 3, pp:66-75, Sept 1991.
- [40] A. D. Wood and J. A. Stankovic, "Denial of Service in Sensor Networks", *IEEE Computer*, vol.35, no.10, pp:54-62, October 2002.
- [41] A. Yaar, A. Perrig and D. Song, "Pi: A Path Identification Mechanism to Defend against DoS Attacks", *IEEE Symposium on Security and Privacy*, pp:93-107, May 2003.
- [42] N. Ye, S. M. Emran, X. Li and Q. Chen, "Statistical Process Control for Computer Intrusion Detection", *DARPA Information Survivability Conference & Exposition II*, vol. 1, pp:3-14, 2001.

- [43] Y. Zhang and W. Lee, "Intrusion Detection in Wireless Ad Hoc Networks", *MOBICOM*, pp:275-283, 2000.
- [44] Y. Zhang, W. Lee and Y. Huang, "Intrusion Detection Techniques for Mobile Wireless Networks", *Mobile Networks and Applications*, pp:1-16, 2002.
- [45] L. Zhou and Z. Haas, "Securing Ad hoc Networks", *IEEE Network*, vol. 13, no. 6, pp:24-30, Nov 1999.

This page intentionally left blank

Chapter 19

WAVEFORM SHAPING TECHNIQUES FOR BANDWIDTH-EFFICIENT DIGITAL MODULATIONS

Hsiao-Hwa Chen

Institute of Communications Engineering

National Sun Yat-Sen University

70 Lien Hai Road, Kaohsiung, Taiwan, ROC

hshwchen@mail.nsysu.edu.tw

Abstract In this chapter, various issues on pulse shaping waveform design for bandwidth efficient digital modulations will be addressed. Here what we concern about the pulse shaping waveforms is the physical appearances or shapes each information data bit or symbol take before carrier modulations. It is well known that the spectral characteristics of a carrier modulated signal can be uniquely determined by its baseband signal before modulation, and thus the waveform shape applied to pre-modulated signals is of ultimate importance to ensure overall bandwidth-efficiency of a wireless communication system. This chapter starts with an overview on the evolution of pulse shaping technologies from early digital modulation schemes to recently emerging new carrier modulations such as quadrature-overlapped modulations. The impact of the pulse shaping technologies on both bandwidth-efficiency and power-efficiency, which are two essential merit parameters of a digital modem, will also be discussed. The majority of the content in this chapter, however, is dedicated to the discussions on shaping pulse design methodology as well as performance analysis of a pulse-shaped digital modem, in which various shaping pulse design methods (including those commonly referred and those recently proposed in the literature) will be presented. To measure the effectiveness of the pulse-shaping techniques, different pulse shaped digital modulations will be evaluated and then compared in terms of their spectral characteristics, the sensitivity of their performance against synchronization timing jitter as well as their bit error rate performance, where the emphasis will be put on the pulse-shaped quadrature-overlapped (QO) modems under bandpass nonlinear channels, which usually exist in satellite and mobile cellular

lar systems. Several new pulse shaped QO modulation schemes, which adopt various novel pulse shaping waveforms generated by time-domain convolution method, will be presented and their performance analysis will be carried out in comparison with other traditional digital modems, such as QPSK, OQPSK, MSK and so forth.

Keywords: Pulse shaping waveform, digital modulation, bandwidth efficiency, power efficiency

1. Introduction

In the recent decay, global wireless revolution has significantly changed the life-style of people around the world. The dream for voice communication anywhere and anytime before has already become the reality in many countries, characterized by the great success of the 2nd generation of mobile cellular networks, in particular the Global System for Mobile (GSM) systems, around the world. It is still a bit hard for everyone to believe that over 250 million people in China today are using mobile phones as the part of their daily life. No one can offer an accurate prediction for the number of wireless services subscribers in China after ten years to come. In Taiwan, the penetration rate for the GSM mobile cellular networks has exceeded 120% at the beginning of 2003, meaning that there are virtually more registered GSM hand phones than the total population of Taiwan. The world has entered a new era of wireless communications since NTT DoCoMo of Japan launched the first 3G commercial service in Tokyo in October 2002, soon followed by another 3G operator based on cdma2000 in the same country. The increasing demand on wideband wireless around the world has already made limited radio spectra an extremely precious resource on the earth. Different from the other resources, such as petroleum, minerals, etc., which might be possible to find some replacement as an energy source, the radio spectra are physically non-replaceable, implying that when it is gone no new wireless services could be added. Unfortunately, the suitable spectra for wireless applications (especially mobile cellular services) span only from a few hundreds MHz to less than a few tens GHz, due to the constraints for radio waves to propagate through the air without being sensitively affected by environmental conditions, such as rain, snow, smoke, dust, etc. It is hardly feasible for a future mobile cellular system to run on a RF carrier higher than millimeter wavelength band, in which the radio signal transmissions could never be made weather-proof, due to the fact that its wavelength is almost in the same dimension of rain-drops, snow-flakes, dust, etc., causing a great penetration loss of the signal. In this sense, the bandwidth efficiency of a wireless system has to be-

come a serious concern in the process of today's communication system architecture design.

The seek for bandwidth efficient digital modulations is not a new subject of interest in the area. In fact, it is one of ever-lasting pursuits in communication field research, which can be traced back to as early as the 60th when Baker and several other scholars first proposed $\pi/4$ -QPSK modulation scheme [1], followed by lots of discussions on practical QPSK-type modulation schemes in the community [2, 3, 4, 5, 15]. In the 70th the study on other modulation schemes, such as FSK [6, 11], OQPSK [7, 9, 10, 11, 12, 15], MSK [8, 9, 10, 11, 12, 13, 15] was also extremely active. Another important bandwidth-efficient digital modulation scheme, GMSK [14, 16], was introduced in the 80th, which had been successfully applied to the popular GSM networks (a 2G mobile cellular standard) around the world today. A great amount of work had been done to address various related issues [12]-[49] in different schemes, such as error rate analysis, bandwidth efficiency comparison under different operational scenarios and so forth.

In general, as far as a digital modulation scheme is concerned, there are several ways that have been commonly adopted to improve its bandwidth efficiency, which include:

- One of the most straight-forward ways is to prolong baseband bit or symbol time-duration as much as possible under the condition that no excessive inter-symbol interference (ISI) will be introduced in transmission. As the overall bandwidth of modulated signal is basically determined by bit or symbol time-duration, the longer each bit or symbol becomes, the narrower the signal bandwidth will occupy.
- Another way to improve the bandwidth efficiency of a digital modem is to minimize the amount of phase fluctuation or shift between adjacent symbols that take different signs or values. In fact, the phase change between two neighboring symbols does not directly affect bandwidth occupancy of the modulated signal. However, when the signal goes through a bandpass hard-limiting element, such as a band-limited nonlinear RF power amplifier and etc., the modulated signal will suffer so-called *side lobe regeneration* problem [7, 8, 15, 39] if the phase of the modulated signal changes abruptly in the boundary of two symbols. Because most power amplifiers always work in a non-linear range to achieve a higher power utilization efficiency, the *side lobe regeneration* problem can be a serious concern.

- Yet another obvious way to achieve a higher bandwidth efficiency is to employ multi-amplitude or/and multi-phase modulations schemes, however at the price of sacrificing power efficiency.
- Finally, we can also use waveform-shaping or pulse-shaping technologies to improve the spectral characteristics of modulated signal due to the fact that bandwidth occupancy of a modulated signal can be uniquely decided by its pre-modulated baseband signal waveform.

There are pros and cons in all aforementioned methods to improve bandwidth efficiency of a digital modulation scheme. Of course, it will be ideal if all methods can be applied jointly though it is hardly possible due to implementation complexity and other practical concerns. The first method has a technical limitation on how to prolong bit or symbol duration without introducing extra ISI. A typical example in using this method to achieve a high bandwidth efficiency is *partial response waveforms*, which usually extend much longer than the width a symbol really needs. A $\sin(\pi t)/\pi t$ pulse has been often taken as one of such partial response waveforms, which span infinitely long in time, and all those pulses will be modulated by different signs to reflect different binary bit information (“+1” or “-1”) and overlapped one by one with each offset by a symbol duration. Theoretically speaking, there will be no ISI if the signal is sampled precisely at the tip of each pulse. However, a great ISI will occur if sampling timing jitter exists. Another very successful example in using this method to improve the bandwidth efficiency is *quadrature overlapped* (QO) modulations, which were first proposed in the 70th by [10], followed by [17] in 1981 and our research work [29]-[33] [35]-[42]. The basic idea of the QO modulations is to use a non-square symbol waveform extended twice as long as its original bit or symbol duration and adjacent extended waveforms are overlapped by half with each other in a quadrature modulator. Therefore, overall signal bandwidth is effectively reduced to only the half if compared with that without waveform extension and overlapping. We should discuss more in detail about the QO modulations in this chapter later.

The second way to improve bandwidth efficiency is reflected on the motivation of the research on continuous phase modulation (CPM) schemes in recent 20 years, starting from QPSK and OQPSK, followed by MSK and then GMSK and so on, with their maximal adjacent-symbol phase shift decreased constantly as the new modulations were introduced, from π (for QPSK), $\pi/2$ (for OQPSK) to zero (for MSK and GMSK). To seek for even smoother phase change, many new CPM schemes were proposed but with their implementation complexity increased too. How-

ever, GMSK can already offer a very robust performance against the side lobe regeneration problem as shown by its very successful application in GSM system. Therefore, it seems at this moment (maybe not in the future) that there is no much room left for the further improvement of bandwidth efficiency if only through improving the phase continuity of modulated signal.

The use of multi-amplitude/phase modulations will seriously affect power-efficiency or detection-efficiency of a receiver. In other words, we improve bandwidth efficiency in this method at the price of a lower power-efficiency. Therefore, this method will never work well if air-link SNR budget is not sufficiently higher than the threshold, which unfortunately is always the case for most communication systems.

Different from the previous three methods, which are limited by many constraints, the waveform shaping technologies, which is also the focal point of this chapter, leave us much more room or degree-of-freedom to improve the bandwidth efficiency of a digital modulation scheme, without introducing too many negative side-effects to the overall performance of a digital communication system. In addition, the implementation of majority waveform shaping technologies is relatively simple and thus they can be used in different digital modulations effectively.

It is noted that a newly emerging multiple access technology, namely ultra-wideband technology (UWB) [43]-[48], needs also proper pulse shaping waveforms in its data modulation and user multiplexing process. In an UWB system data will be sent via extremely narrow pulses, whose width can be as narrow as a fraction of an nano-second. Obviously, the pulse shapes have a lot to do with the system bandwidth efficiency, which is of ultimate interest to us. In this sense, the discussions about pulse shaping technologies in this chapter can also be found relevant to the UWB system design.

The rest of the chapter are organized as follows. Section 2 will address the issues on the roles the pulse shaping techniques play under the context of bandwidth efficient digital modulations. Section 3 discusses various methodologies used to design pulse shaping waveforms, including intuitive methods, piece-wise fitting methods and time-domain convolution methods. The merit parameters of pulse shaping waveforms (PSW's) and their major performance measures in a digital modulation scheme using designed PSW's are introduced in section 4, followed by the illustrations on how to evaluate spectral efficiency of a pulse-shaped digital modulation in bandpass nonlinear channels in section 5. In section 6, the issues on robustness of a pulse-shaped digital modem against sampling timing jitter will be discussed. Finally, we will give a summary of the chapter.

2. Roles of Waveform Shaping Techniques

The evolution of digital modulations from BPSK to all its later derivatives can be traced in fact from a perspective of waveform-shaping technology. Table (19.1) lists several widely used digital modulation schemes together with their baseband signal waveform shapes and their major spectral characteristics. It is noted from Table (19.1) that, except BPSK, all listed modulation schemes are based on quadrature modulation with I and Q two channels in their modem hardware implementation. Under such a very similar structure, they differ from one another mainly from their different waveform shapes in their baseband signal before carrier modulation process. A generic model for any quadrature modulations is shown in Figure (19.1), where *pulse shaping filter* can be replaced by any particular digital filter to fit for a particular quadrature modulation. For instance, if the output from the filter is a $2T$ -long square waveform, then Figure (19.1) shows a OQPSK transmitter. On the other hand, if the output from the filter becomes *half-cycle sinusoidal* waveforms, then the figure gives a MSK modulator. If we replace the filter with a Gaussian waveform generator, Figure (19.1) in fact means the block diagram of a GMSK transmitter. Please also note that the *Post-shaping processor* units, which are reserved for quadrature-overlapped modulations [17, 24, 25, 27], should not function for conventional quadrature modulations, such as OQPSK, MSK and GMSK.

Table 19.1. Several popular digital modulation schemes with their baseband waveform shapes and other spectral characteristics.

| | BPSK | QPSK | OQPSK | MSK | GMSK | QORC ^a |
|-------------------|-------------|-------------|-------------|-----------------|-------------|-------------------|
| Waveform | Square | Square | Square | Half-sinusoidal | Gaussian | Raised cosine |
| MHC ^b | I | I&Q | I&Q | I&Q | I&Q | I&Q |
| SWOD ^c | No | No | Yes | Yes | Yes | Yes |
| SWEO ^d | No | No | No | No | No | Yes |
| MPS ^e | π | π | $\pi/2$ | 0 | 0 | 0 |
| FNBM ^f | fT | fT | fT | $1.5 fT$ | $1.5 fT$ | fT |
| SFOR ^g | $(fT)^{-2}$ | $(fT)^{-2}$ | $(fT)^{-2}$ | $(fT)^{-4}$ | $(fT)^{-5}$ | $(fT)^{-6}$ |

^a QORC: Quadrature overlapped raised cosine modulation [17].

^b MHC: Modem hardware configuration.

^c SWOD: IQ channel Symbol waveform offset delayed.

^d SWEO: Symbol waveform extension and overlap.

^e MPS: Maximum phase shift at the boundary of two adjacent symbols.

^f FNBM: First null beside the main lobe in unit of fT , where f is the frequency and T is the pre-modulated bit duration.

^g SFOR: Spectral fall-off rate.

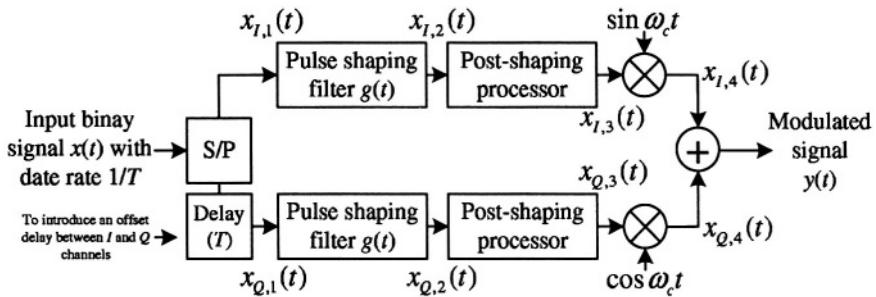


Figure 19.1. A generic quadrature modulator, where *Pulse shaping filter* unit is for particular waveform shaping, *Post-shaping processor* unit is for quadrature-overlapped modulation schemes [17].

It is seen clearly from Table (19.1) that the evolution of bandwidth efficient digital modulations is in fact closely associated with the development of waveform shaping technologies. Then, why the waveform shaping technology can play so much an important role in determining the overall bandwidth efficiency of a digital modulation scheme? This can be explained in the text followed with the same generic model of a digital modulation scheme shown in Figure (19.1) as follows.

From Figure (19.1), we have

$$x_{I,2}(t) = x_{I,1}(t) \otimes g(t) \quad (19.1)$$

$$x_{Q,2}(t) = x_{Q,1}(t) \otimes g(t) \quad (19.2)$$

where the notation $x(t) \otimes y(t)$ stands for convolution operation between two time-domain functions $x(t)$ and $y(t)$, and $g(t)$ is the impulse response of the pulse-shaping filter. If ignoring the *Post-shaping processor* unit in Figure (19.1) for illustration clarity, we have

$$y(t) = x_{I,4}(t) + x_{Q,4}(t) \quad (19.3)$$

$$= x_{I,2}(t) \sin \omega_c t + x_{Q,2}(t) \cos \omega_c t \quad (19.4)$$

$$= (x_{I,1}(t) \otimes g(t)) \sin \omega_c t + (x_{Q,1}(t) \otimes g(t)) \cos \omega_c t \quad (19.5)$$

Therefore, if defining $y(t) \Leftrightarrow Y(f)$, $g(t) \Leftrightarrow G(f)$, $x_{I,1}(t) \Leftrightarrow X_{I,1}(f)$ and $x_{Q,1}(t) \Leftrightarrow X_{Q,1}(f)$ are Fourier transform pairs, we have

$$\begin{aligned} & Y(f) \\ &= j \frac{1}{2} (X_{I,1}(f + f_c)G(f + f_c) - X_{I,1}(f - f_c)G(f - f_c)) \\ &\quad + \frac{1}{2} (X_{Q,1}(f - f_c)G(f - f_c) + X_{Q,1}(f + f_c)G(f + f_c)) \end{aligned} \quad (19.6)$$

Thus, the power spectral density (PSD) of output modulated signal is

$$\begin{aligned}
 & |Y(f)|^2 \\
 = & |j\frac{1}{2}(X_{I,1}(f + f_c)G(f + f_c) - X_{I,1}(f - f_c)G(f - f_c)) \\
 & + \frac{1}{2}(X_{Q,1}(f - f_c)G(f - f_c) + X_{Q,1}(f + f_c)G(f + f_c))|^2 \quad (19.7)
 \end{aligned}$$

from which it can be seen that the PSD of the output signal is uniquely determined by the Fourier transform of the pulse shape or $G(f)$ as either $X_{I,1}(f)$ or $X_{Q,1}(f)$ is Fourier transform of a random data sequence which is non-alterable component. Therefore, if we can find a time waveform of $g(t)$ whose Fourier transform $G(f)$ has a very concentrated spectral shape around zero frequency, then we will also have a output modulated signal with a very compact form around the carrier frequency f_c , giving rise to a promising bandwidth efficiency. Please take note that this conclusion is valid for either single channel modulation schemes (such as BPSK) or quadrature modulations with I and Q channels (such as QPSK, etc.). Thus, the key for us to find a bandwidth efficient digital modulation scheme is equivalently to search for a promising time waveform with a very narrow PSD function. This is the most important fact that should be properly underlined throughout this chapter.

The great importance of waveform shaping technologies in the development of digital modulations can also be well explained in a more descriptive way in the sequel. Let us again look at Figure (19.1), which can be viewed as a generic model for many commonly used digital modulators. If only one channel is present, such as BPSK, we could take only I -channel into account without losing generality. Again, let us omit the block of *Post-shaping processor* for illustration clarity. Now we can follow the signal flow from the left hand side to the right hand side in the block diagram. It can be seen from the figure that all signal processing stages, except for the convolution operation taking place in *Pulse-shaping filter* unit, will never alter the spectral characteristics of the input signal. For example, the multiplication of the carrier, either $\sin\omega_ct$ or $\cos\omega_ct$, with the signal $x_{I,3}(t)$ or $x_{Q,3}(t)$ will not change its original spectral shape at all, only shifting its spectral center from zero to the carrier frequency ω_c . Also, the summation operation at the joint point of I and Q channels will not affect the spectral characteristics of the signals either. In this way, we can easily conclude from here that the single most important part of the modulator that determines the spectral characteristics of the output signal is the impulse response of *Pulse-shaping filter* $g(t)$.

Obviously, the waveform shapes can also sensitively affect many other merit parameters of a digital modulation scheme, as they do with the bandwidth efficiency. For instance, the waveform shapes have a great impact on the opening of the eye-diagrams of a digital modulation scheme. They can also affect the immunity of a digital modem against timing jitter resulted from unstable symbol or bit synchronization circuit. We will address all those issues in the later part of this chapter.

We have to admit that all traditional digital modulations have chosen to use relatively simple bit waveforms and thus little improvement can be achieved in their spectral characteristics as well as other merit parameters, such as power-efficiency and some other time-domain characteristics, such as the robustness against sampling timing jitter, etc. The pulse shapes found in those conventional modulation schemes include square (used by BPSK, QPSK, OQPSK), half-cycle sinusoidal (used by MSK), Gaussian (used in GMSK) and raised cosine (used in QORC [17]) waveforms, etc., which are intuitive or primitive in nature. The correspondence between symbol waveform shapes and overall performance of a modulated signal in both time- and frequency-domains motivates us to search for much more deliberately designed waveform shapes. It is indeed a fascinating research topic worth our further study.

It is not surprising at all that a carefully designed pulse shaping waveform can be used for any type of digital modulation schemes with immediate improvement in both time- and frequency-domains performance. Let us take the simplest modulation, BPSK, as an example. Replacement of square waveform with raised-cosine pulse in BPSK modulator can directly improve the bandwidth efficiency of BPSK modulated signal, giving rise to a narrower main lobe and faster spectral roll-off rate than those before the replacement.

To have a more objective view on impact of waveform shaping techniques on the overall performance of a digital modem, let us look at some fundamental requirements for a desirable modulation scheme. The following considerations should be taken into account whenever design of a digital modulation scheme is concerned.

- First of all, the bandwidth efficiency of a modem is of primary concern to us. To have an objective measurement of the bandwidth efficiency, there are many different merit parameters which might be used, such as the width of main lobe of power spectral density (PSD) function of a modulated signal, the out-of-band emission defined as the amount of power left outside a specific bandwidth, the spectral roll-off rate defined as the reducing rapidity in PSD side lobe levels with respect to the normalized offset frequency from the carrier fT , etc.

- Another important parameter for a digital modem is its power efficiency, which sometimes can also be called detection efficiency. The power efficiency can be best described by the relation between bit error rate and signal-to-noise-ratio (SNR), which exactly means how efficiently a modem can utilize its initially transmitted power in recovering the information data at a receiver. It is clear that the power efficiency and bandwidth efficiency is a dual, which usually counteract with each other. In other words, it is difficult for a modem to achieve good performance of the both at the same time. A simple example to illustrate this phenomenon is that the use of a multi-amplitude/phase modulation can effectively improve the bandwidth efficiency, but at the price of decreasing the power efficiency at the same time.
- Complexity is another important concern to a digital modulation scheme, which should include both transmitter and receiver hardware and software complexity. It should be stressed that a good modulation scheme does not necessarily mean a high complexity. Our objective is to design a well-performing modulation scheme with a reasonable complexity. However, it should also be pointed out that the *complexity* is only a time-dependent term, which has different measures with the time elapses. It is easy to understand that a circuit that was considered as *very complex* ten years ago is not necessarily the same today due to the fast advancement of microelectronics industry. Therefore, in this sense, the importance of the consideration to the complexity of a digital modem should never be over exaggerated if compared with that for bandwidth efficiency and power efficiency of a modem.
- Yet another concern for a digital modulation is its performance under some particular working environment, such as band-limited non-linear channels, fading channels, multipath channels, noisy channels and other interfering channels, etc.
- When searching for a good digital modulation scheme, we should also pay attention to its performance under sampling time jitter, which is associated with the behaviors of synchronization acquisition and tracking systems at a receiver. This is also a very important issue whenever practical applications will be concerned.

With all aforementioned considerations in mind, we will be ready to discuss the issues on waveform design for well-performing digital modulation schemes. In doing so, we should pay a great attention to the bandwidth efficiency of the schemes in this chapter, although the other

performance parameters of the schemes can also be the subject of our interest whenever necessary.

3. Waveform Design Methodologies

To design promising pulse shaping waveforms to improve the performance of a digital modulation is a traditional research topic, as being evident in the reference list related to the subject and given at the end of this chapter. Some interesting waveforms as well as their generation methods were proposed in the literature [50]-[65], some of which can be summarized as follows.

3.1 Intuitive Methods

The majority of the pulse shaping waveforms used in early digital modulations were introduced in a non-systematic approach or by so called *intuitive methods*, which will be discussed in this section.

The pulse shapes generated using those intuitive methods include half-cycle sinusoidal waveform used in MSK, Gaussian functions for GMSK [14, 16], raised-cosine waveform used in QORC [17], cosine-roll off waveforms and $\text{sinc}(t)$ functions, etc, where all of which are finite-time pulses except for Gaussian waveforms and $\text{sinc}(t)$ functions defined as $\sin(\pi t)/\pi t$. The later has been found in the applications of some partial response transmission systems.

In the early days MSK modulation was often viewed as a special case of continuous phase FSK (CPFSK) modulation due to the fact that its phase change has been made continuous as a very important step forward from OQPSK modulation, whose phase change still abruptly though at a relatively small value $\pi/2$ if compared to π in QPSK. However, from the view point of hardware implementation MSK can also be considered as a special case of OQPSK whose square bit pulse waveforms in both I and Q channels are replaced by half-cycle sinusoidal waveforms. In this sense, the block diagram of a MSK modulator can be shown in Figure (19.2), which in fact is a modified version of Figure (19.1) with only *Post-shaping processor* unit omitted.

The baseband symbol waveform in both I and Q channels in a MSK modulator is shown in Figure (19.3), where it has been assumed that all symbols carry the same sign +1 for representation clarity.

The baseband waveform of a MSK modulator shown in Figure (19.3) will completely determine the spectral characteristic features of MSK modulated signal, which include

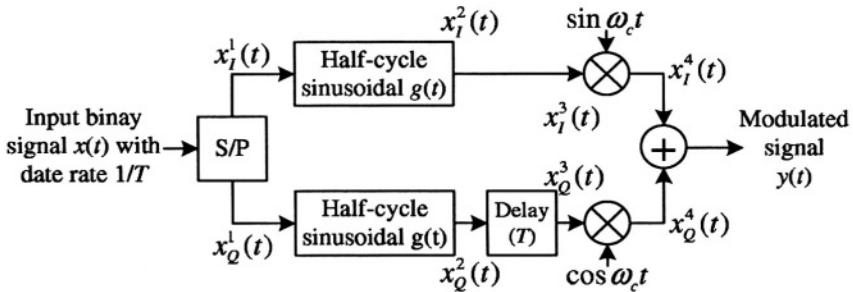


Figure 19.2. Block diagram of MSK modulator presented from waveform shaping perspective.

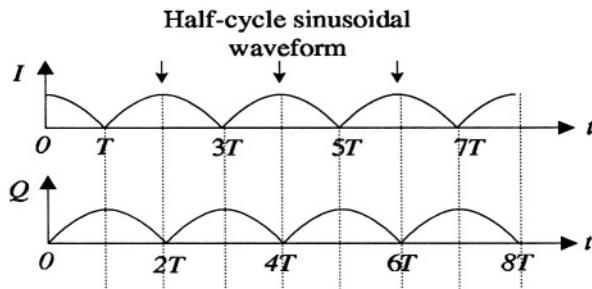


Figure 19.3. Baseband symbol waveform in both I and Q channels in a MSK modulator.

- continuous phase change with time in the modulated signal, which can effectively reduce side lobe regeneration effect in a band-limited non-linear channel;
- enhanced power concentration in the PSD main lobe;
- a fast spectrum fall off at a rate proportional to f^{-4} for large values of f . In contrast, the offset QPSK spectrum falls off at a rate proportional to only f^{-2} , as a direct benefit from using half-cycle sinusoidal waveforms instead of square pulse in OQPSK.
- a truly constant envelope in MSK modulated signal in the time domain, which is important to reduce the distortion in a hard-limiting channel.

Thus, the advantages gained by the transformation from OQPSK to MSK clearly shows true power of waveform shaping technology.

GMSK [14, 16] is another vivid example of bandwidth efficiency enhancement by using waveform shaping techniques. The GMSK modulator structure is very similar to that of MSK, with only difference in the

replacement of *half-cycle sinusoidal* pulse shaping filter with *Gaussian* pulse shaping filter in Figure (19.2). However, it has to be mentioned that the Gaussian waveform is not a finite time function, a big difference from the *half-cycle sinusoidal* waveform used in MSK modulator. The impulse response of a Gaussian pulse shaping filter can be written as

$$g_{gaussian}(t) = \frac{\sqrt{\pi}}{\alpha} e^{-\frac{\pi^2 t^2}{\alpha^2}} \quad (19.8)$$

The transfer function or the Fourier transform of (19.8) is

$$G_{gaussian}(f) = e^{-\alpha^2 f^2} \quad (19.9)$$

Therefore, there is an important parameter α in a Gaussian waveform, which controls its waveform extension in the time domain, as shown in Figure (19.4). In (19.8) and (19.9), the parameter α is closely related to 3-dB bandwidth (B_{3dB}) of the Gaussian waveform pulse shaping filter, whose relation can be represented by

$$\alpha = \frac{\sqrt{2 \ln 2}}{B_{3dB}} = \frac{1.1774}{B_{3dB}} \quad (19.10)$$

Obviously, the 3-dB bandwidth of the filter will shrink or the Gaussian waveform shape becomes extended as α increases. GMSK modulated signal has a very good spectral characteristics, which can be even better than that of MSK, especially when it is applied to a system with non-linear power amplifier, which is often used in a mobile handset. Similar to MSK, GMSK modulated signal has also a constant signal envelope and zero phase shift between adjacent symbols. In addition, the side lobes of GMSK PSD function fall off even faster than those of MSK. Therefore, GMSK modulation has been successfully used in GSM system, which is the most popular cellular mobile standard in the world today.

However, it has to be mentioned that GMSK waveform is not a finite time function, as shown in Figure (19.4). Therefore, choosing a wrong (too large) α value will result in serious ISI, greatly impairing detection efficiency at a receiver. On the other hand, as mentioned earlier, a large α helps reduce the overall bandwidth occupancy of modulated signal. Therefore, a careful trade-off should be made when deciding the values of α in GMSK modulation.

Raised-cosine waveform is another widely used pulse shape in digital modulations. The most important example of its application is in QORC modulation, which was first proposed by Mark C. Austin and Ming U. Chang in 1981 [17]. The abbreviation *QORC* means *quadrature-overlapped raised cosine*, which uses raised-cosine waveform as its pulse

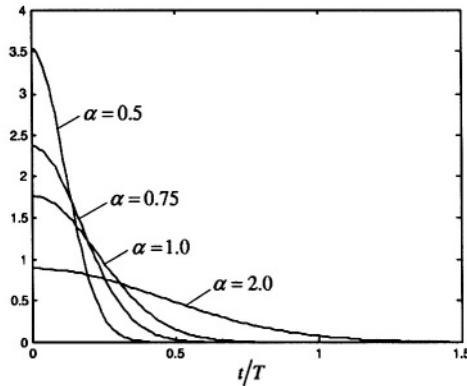


Figure 19.4. Gaussian waveform with different α values used in GMSK modulator, where T is the symbol duration.

shape and quadrature-overlapped scheme as its modulation structure, as shown in Figure (19.1). A $4T$ duration raised-cosine waveform can be represented by

$$g_{rc}(t) = \frac{1}{2}(1 - \cos \frac{\pi t}{2T}), \quad (0 < t \leq 4T) \quad (19.11)$$

where T stands for the symbol duration of the input signal to the modulator, as shown in Figure (19.1). The PSD function of the raised-cosine waveform can be shown as

$$\begin{aligned} \Phi_{rc}(f) = |G_{rc}(f)|^2 &= 4T^2 \left[2 \frac{\sin 4\pi fT}{4\pi fT} + \frac{\sin 4\pi(4fT + 1)}{\pi(4fT + 1)} \right. \\ &\quad \left. + \frac{\sin 4\pi(4fT - 1)}{\pi(4fT - 1)} \right]^2 \end{aligned} \quad (19.12)$$

In fact, the hardware implementation of a QORC modulator is quite similar to a MSK modulator. The main difference lies on its pulse waveform *overlapping*, in addition to their different pulse shapes used: QORC uses raised-cosine and MSK uses half-cycle sinusoidal waveforms. Therefore, the generic model for a digital modulation scheme shown in Figure (19.1) is still applicable to QORC modulation with the *post-shaping processor* unit taking care of waveform overlapping action. To illustrate this point more clearly, we have drawn the baseband symbol waveforms for both MSK and QORC modulations in Figure (19.5), where the time-domain waveforms for QORC appear as their individual symbols (assume to be all +1), instead of the sum of the overlapped waveforms, for illustration clarity.

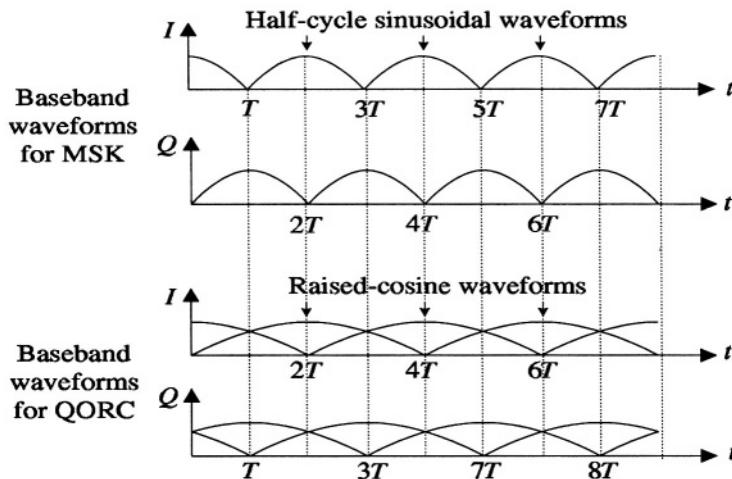


Figure 19.5. Comparison of baseband waveforms of MSK and QORC modulations, where it is assumed that all symbols carry the same information "+1" for illustration clarity.

The QO modulation is of great interest to us in this chapter due to the following reasons:

- The QO modulations can use twice as long symbol waveforms as those in conventional quadrature modulation schemes (such as MSK, GMSK and etc.) to enhance its overall bandwidth efficiency (occupying only about half of the bandwidth required by MSK, GMSK and etc.) without reducing the data transmission rate, as shown in Figure (19.5).
- Although waveform overlapping is used in QO modulations, there should be no ISI if each symbol can be sampled at right the center of the waveform. Therefore, the way used by QO modulation to improve bandwidth efficiency is much more effective than other schemes that also try to use prolonged symbol waveforms to improve the bandwidth efficiency, such as the case in a partial response system.
- The QO modulations retain a relatively simple implementation hardware if compared with many other digital modulation schemes with a similar bandwidth efficiency. The modulator structure is shown in Figure (19.1), which is almost the same as a conventional OQPSK modulator.

As far as bandwidth efficiency is concerned, QORC outperforms MSK with a comfortable margin. Let us take normalized PSD level of -20 dB as a threshold. MSK needs a bandwidth of about $fT = 0.67$ to reach the level; while QORC requires only $fT = 0.41$ to achieve the same level.

However, the QO modulation has also its technical limitations, which can be summarized as follows:

- Usually QO modulated signal is not a constant-envelope signal. The fluctuation in the envelope of QO modulated signal will have some negative impact on its performance in a non-linear channel. Let us define a parameter ρ to measure the severity of such fluctuation as the ratio between the maximum and minimum envelope amplitudes of the modulated signal. QORC gives about $\rho = 0.707$.
- Another concerned issue with a QO modulation is its sensitivity against possible sampling timing jitter. As shown in Figure (19.5), QORC will not produce any ISI only if each symbol is sampled at exactly the center of the symbol. Otherwise, the ISI is inevitable.

Fortunately, the above two major problems with QO modulations can be significantly improved or even eliminated if we can use some other more appropriate pulse shaping waveforms instead of the *raised-cosine* function. Thus, the usefulness of the QO modulations will definitely offer us some more degrees-of-freedom in designing a proper digital modulation scheme for some particular applications. This will be the issue to be discussed in much a great detail in the text followed.

3.2 Piece-Wise Fitting Methods

The piece-wise fitting method was first introduced by D. S-Dias and K. Feher in [20, 24] and used for baseband pulse shaping in $\pi/4$ -DQPSK modulation to improve the spectral efficiency of a transceiver with a non-linear amplifier in land mobile environment. It was also demonstrated in [24] that with proposed new pulse shaping method, which resulted in $\pi/4$ -QORC (Quadrature Overlapped Raised-Cosine) and $\pi/4$ -CTPSK (Controlled Transmission PSK) modulations, the spectral side lobes regeneration after nonlinear amplification can be reduced significantly. Thus, the proposed pulse shaping techniques allow closer channel spacing if compared to a conventional $\pi/4$ -DQPSK, the standard modem scheme for the US digital mobile communication system.

The piece-wise fitted pulse shaping technique used in [24] employed the following two-segmented functions:

If $a_n = \pm 1$, then

$$i_1(t) = \begin{cases} a_{n-1}, & nT < t \leq (n + \frac{1}{4})T \\ \frac{a_{n-1}+a_n}{2} - \frac{a_n-a_{n-1}}{2} \cos \frac{4\pi}{3}(\frac{t}{T} - \frac{1}{4}), & (n + \frac{1}{4})T < t \leq (n + 1)T \end{cases} \quad (19.13)$$

$$q_1(t) = \begin{cases} \frac{b_{n-1}+b_n}{2} - \frac{b_n-b_{n-1}}{2} \cos \frac{4\pi}{3}\frac{t}{T}, & nT < t \leq (n + \frac{3}{4})T \\ b_n, & (n + \frac{3}{4})T < t \leq (n + 1)T \end{cases} \quad (19.14)$$

If $a_n = 0$, then

$$i_2(t) = \begin{cases} \frac{a_{n-1}+a_n}{2} - \frac{a_n-a_{n-1}}{2} \cos \frac{4\pi}{3}\frac{t}{T}, & nT < t \leq (n + \frac{3}{4})T \\ a_n, & (n + \frac{3}{4})T < t \leq (n + 1)T \end{cases} \quad (19.15)$$

$$q_2(t) = \begin{cases} b_{n-1}, & nT < t \leq (n + \frac{1}{4})T \\ \frac{b_{n-1}+b_n}{2} - \frac{b_n-b_{n-1}}{2} \cos \frac{4\pi}{3}(\frac{t}{T} - \frac{1}{4}), & (n + \frac{1}{4})T < t \leq (n + 1)T \end{cases} \quad (19.16)$$

If $a_n = \pm \frac{1}{\sqrt{2}}$ and $|a_n - a_{n-1}| > |b_n - b_{n-1}|$, then

$$i_3(t) = i_2(t) \quad (19.17)$$

$$q_3(t) = q_2(t) \quad (19.18)$$

If $a_n = \pm \frac{1}{\sqrt{2}}$ and $|a_n - a_{n-1}| < |b_n - b_{n-1}|$, then

$$i_4(t) = i_1(t) \quad (19.19)$$

$$q_4(t) = q_1(t) \quad (19.20)$$

where a and b stand for the $\pi/4$ -mapped signal levels at the input to a baseband signal processing unit (which is equivalent to the combination of *Pulse-shaping filter* and *Post-shaping processor* units in Figure (19.1)) in the I and Q channels respectively. The subscripts n and $n-1$ represent the n th and the $(n-1)$ th symbol intervals respectively. The processed output signals in the I and Q channels are given by $i(t)$ and $q(t)$ respectively.

Please note that the above piece-wise fitted pulse shaping technique carries two salient features:

- It is a single-interval pulse shaping technique with no overlapping between adjacent symbols; while the raised-cosine pulse shaping technique used in QORC [17] allows half-symbol overlapping between two adjacent symbols.

- The pulse shaping processes happened to the I and Q channels are not carried out independently; while those in the QORC can be done independently.

Another example of using piece-wise fitting method to implement digital modulation is the two new QO modulations proposed earlier by us in [38], namely *Quadrature-Overlapped Triangular Cosine* (QOTC) modulation and *Modified Quadrature-Overlapped Triangular Cosine* (MQOTC) modulation.

Although these two QO modulations use also the piece-wise fitting technique in their pulse shaping procedure prior to the carrier modulation, as does in the previous case [20, 24], their working principle is quite similar to QORC modulation, which uses raised cosine function as its pulse shaping waveform (PSW). The most important aspect of them is their unique PSWs employed in their modulators. QOTC modulation uses a combination of a triangular and a cosine function as its PSW; and MQOTC modulation uses the second order version of triangular and cosine functions as its PSW.

Therefore, when compared with QORC modulation, which uses simple raised cosine function as its PSW, we can see the delicacy in these new modulations and their unique PSWs that result in much improved bandwidth efficiency and power efficiency. At the same time, the modulator structure of these two new modulations are almost exactly the same as that of QORC modulation, except for their different pulse shaping filters.

The time domain impulse response of the pulse-shaping filter used by QOTC modulation is given by

$$g_{qotc}(t) = \begin{cases} \frac{t}{2T} - \frac{1}{2\pi} \sin \frac{\pi t}{T}, & 0 < t \leq 2T \\ 2 - \frac{t}{2T} + \frac{1}{2\pi} \sin \frac{\pi t}{T}, & 2T < t \leq 4T \end{cases} \quad (19.21)$$

where T is the bit duration of input signal to the QOTC modulator, whose hardware structure is similar to that shown in Figure (19.1) that should use equation (19.21) as the impulse response of its *Pulse-shaping filter*.

It can be clearly seen that equation (19.21) in fact can be decomposed into two parts: one part is a triangular waveform as

$$g_{qotc,1}(t) = \begin{cases} \frac{t}{2T}, & 0 < t \leq 2T \\ 2 - \frac{t}{2T}, & 2T < t \leq 4T \end{cases} \quad (19.22)$$

and the other part is two back-to-back sinusoidal waveforms whose representation is

$$g_{qotc,2}(t) = \begin{cases} -\frac{1}{2\pi} \sin \frac{\pi t}{T}, & 0 < t \leq 2T \\ \frac{1}{2\pi} \sin \frac{\pi t}{T}, & 2T < t \leq 4T \end{cases} \quad (19.23)$$

Therefore, we have

$$g_{qotc}(t) = g_{qotc,1}(t) + g_{qotc,2}(t) \quad (19.24)$$

The QOTC time-domain waveform itself is shown in Figure (19.6), where we can see how the triangular pulse and two back-to-back sinusoidal waveforms are combined to form a complete QOTC PSW.

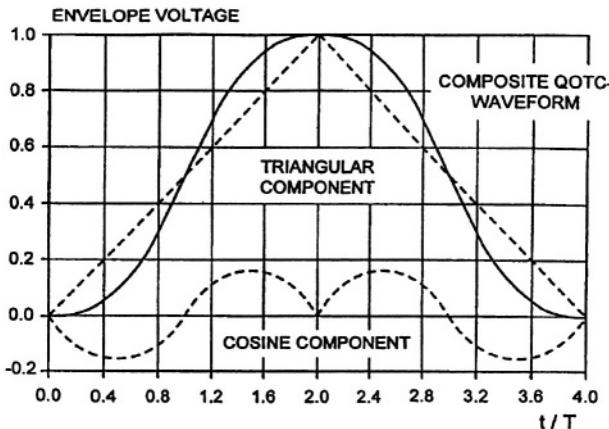


Figure 19.6. QOTC time-domain waveform, which can be decomposed into two components: a triangular pulse and two back-to-back sinusoidal waveforms, where T is the bit duration of input digital signal to the QOTC modem.

Similarly, we have the time-domain impulse response of the MQOTC pulse shaping filter as

$$g_{mqotc}(t) = \begin{cases} \frac{t^2}{2T^2} - \frac{1}{2\pi^2} \sin^2 \frac{\pi t}{T}, & 0 < t \leq T \\ \frac{2t}{T} - \frac{t^2}{2T^2} - 1 + \frac{1}{2\pi^2} \sin^2 \frac{\pi t}{T}, & T < t \leq 3T \\ -\frac{4t}{T} + \frac{t^2}{2T^2} + 8 - \frac{1}{2\pi^2} \sin^2 \frac{\pi t}{T}, & 3T < t \leq 4T \end{cases} \quad (19.25)$$

from which it is observed that, different from the QOTC modulation, the PSW of MQOTC modulation consists of the linear term, the 2nd order non-linear terms and 2nd order sinusoidal terms in its three-segment representation as shown in (19.25), resulting in a more complex equation

if compared to that of QOTC. Therefore, in this sense, the MQOTC modulation can be viewed as a *2nd order modified* version of QOTC waveform.

The PSW of MQOTC can also be decomposed into two major parts, one being the rational functions part and the other being the 2nd order trigonometric functions, shown as follows:

$$g_{mqotc,1}(t) = \begin{cases} \frac{t^2}{2T^2}, & 0 < t \leq T \\ \frac{2t}{T} - \frac{t^2}{2T^2} - 1, & T < t \leq 3T \\ -\frac{4t}{T} + \frac{t^2}{2T^2} + 8, & 3T < t \leq 4T \end{cases} \quad (19.26)$$

$$g_{mqotc,2}(t) = \begin{cases} -\frac{1}{2\pi^2} \sin^2 \frac{\pi t}{T}, & 0 < t \leq T \\ \frac{1}{2\pi^2} \sin^2 \frac{\pi t}{T}, & T < t \leq 3T \\ -\frac{1}{2\pi^2} \sin^2 \frac{\pi t}{T}, & 3T < t \leq 4T \end{cases} \quad (19.27)$$

Therefore we also have

$$g_{mqotc}(t) = g_{mqotc,1}(t) + g_{mqotc,2}(t) \quad (19.28)$$

The time-domain PSW of MQOTC is shown in Figure (19.7)

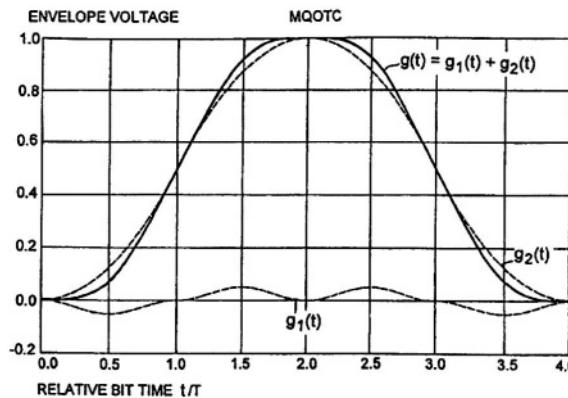


Figure 19.7. MQOTC time-domain waveform, which can be decomposed into two components: 2nd order trigonometric functions and the 1st and 2nd order rational functions, where T is the bit duration of input digital signal to the MQOTC modem.

As the input signal to the pulse-shaping filters in Figure (19.1) is a stream of square bit pulses, it can be shown that the transfer functions for QOTC and MQOTC pulse-shaping filters $g_{qotc}(t)$ and $g_{mqotc}(t)$ should take the forms of

$$G_{qotc}(f) = \frac{\sin 2\pi fT}{2\pi fT} \frac{e^{-j2\pi fT}}{(1 - 2fT)} \quad (19.29)$$

and

$$G_{mqotc}(f) = \left(\frac{\sin \pi fT}{\pi fT}\right)^2 \frac{e^{-j2\pi fT}}{1 - (fT)^2} \quad (19.30)$$

respectively, in order to obtain the correct PSW's at the output of the filters. Again, we can also see the similarity between (19.29) and (19.30).

It can be shown that the PSD's for the QOTC and MQOTC modulated signals can be written as

$$\Phi_{qotc}(f) = \left(\frac{\sin 2\pi fT}{2\pi fT}\right)^4 \frac{1}{(1 - (2fT)^2)^2} \quad (19.31)$$

and

$$\Phi_{mqotc}(f) = \left(\frac{\sin 2\pi fT}{2\pi fT}\right)^2 \frac{1}{(1 - (fT)^2)^2} \left(\frac{\sin \pi fT}{\pi fT}\right)^4 \quad (19.32)$$

respectively, which have quite similar expressions. In Figures (19.8) and (19.9) the PSD's of QOTC and MQOTC modulations are compared with that of MSK.

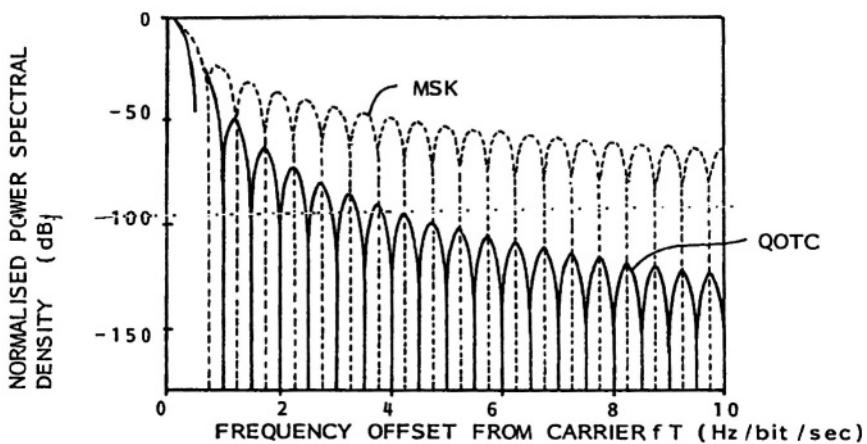


Figure 19.8. PSD function of QOTC in comparison with that of MSK.

It can be seen from Figures (19.8) and (19.9) that the QOTC and MQOTC modulations have the advantages of fast roll-off PSD side lobes and concentrated power nearby the main lobe over MSK. The PSD's of QOTC and MQOTC have a side lobe roll-off rate of $(fT)^{-8}$ and $(fT)^{-10}$, respectively, when compared with that of MSK being $(fT)^{-4}$.

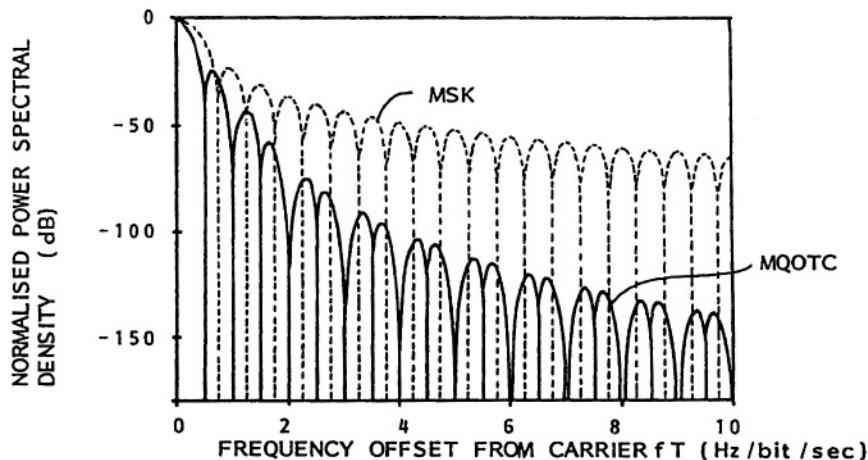


Figure 19.9. PSD function of MQOTC in comparison with that of MSK.

In addition, we can see that the MQOTC has a PSD main lobe of $0.5fT$, which is narrower than that ($0.75fT$) of MSK. Therefore, it is a quite promising bandwidth efficient digital modulation scheme.

However, we have to admit that the piece-wise fitting method is still not a systematic approach to design suitable PSWs for bandwidth efficient digital modulations. They rely on try-and-error in the whole design process and a lot of time is required to find a suitable PSW for a particular application.

3.3 Time-Domain Convolution Methods

The time-domain convolution (TDC) method is a systematic approach proposed by us, which can be used to generate various promising PSW's to suit for different applications in bandwidth efficient digital modems.

Assume that there are two time-domain functions, $f_1(t)$ and $f_2(t)$, and their Fourier transform pairs, $f_1(t) \Leftrightarrow F_1(f)$ and $f_2(t) \Leftrightarrow F_2(f)$. Letting $f(t) \Leftrightarrow F(f)$, we have

$$f(t) = f_1(t) \otimes f_2(t) \Leftrightarrow F(f) = F_1(f)F_2(f) \quad (19.33)$$

where the operator " \otimes " stands for the time-domain convolution between $f_1(t)$ and $f_2(t)$. Thus, $|F_1(f)|^2$ and $|F_2(f)|^2$ are normalized PSD for $f_1(t)$ and $f_2(t)$, respectively.

The principle of the TDC method follows the two simple observations, which can be explained in the sequel.

- The normalized PSD of resultant signal $f(t)$ (yielded from time-domain convolution between $f_1(t)$ and $f_2(t)$), $|F(f)|^2$, always has a faster side lobe roll-off rate than that of either $|F_1(f)|^2$ or $|F_2(f)|^2$. Therefore, the $|F(f)|^2$ usually possesses a better bandwidth efficiency than that of either $|F_1(f)|^2$ or $|F_2(f)|^2$.
- If $f_1(t)$ and $f_2(t)$ are two equal length time pulses, the output pulse $f(t)$ from convolution between $f_1(t)$ and $f_2(t)$ has always a duration about twice as long as that of either $f_1(t)$ or $f_2(t)$.

The above two observations in fact are closely related in a sense that a prolonged time waveform can effectively yield a relatively concentrated power distribution in frequency-domain. However, using a prolonged waveform while keeping the data rate unchanged can only be made feasible through application of the QO modulations. This explains the reason why majority of the new PSW's proposed in this chapter are used in QO modulations, not others.

Based on the above two observations, we can generate many interesting PSW's for the applications in QO-based bandwidth efficient digital modulations. Let us first start with conventional MSK modulation, whose PSW is a $2T$ -long half-cycle sinusoidal waveform, which can be written as

$$g_{msk}(t) = \sin \frac{\pi t}{2T} \quad (0 < t \leq 2T) \quad (19.34)$$

where T is the bit duration of input digital signal into the MSK modem.

The normalized output from convolution of two identical $g_{msk}(t)$ becomes

$$g_{qocc}(t) = \begin{cases} \frac{1}{\pi} \sin \frac{\pi t}{2T} - \frac{t}{2T} \cos \frac{\pi t}{2T}, & 0 < t \leq 2T \\ (\frac{t}{2T} - 2) \cos \frac{\pi t}{2T} - \frac{1}{\pi} \sin \frac{\pi t}{2T}, & 2T < t \leq 4T \end{cases} \quad (19.35)$$

whose duration has been extended into $4T$, which is twice as long as that in (19.34). We have employed this PSW in a QO modulation we named as quadrature-overlapped convoluted cosine (QOCC) modulation [38]. The QOCC PSW has been shown in Figure (19.10).

The QOCC PSW can also be viewed as the combination of two waveforms, as shown in Figure (19.10). The normalized PSD of QOCC modulated signal, which is the product of PSD's of two identical half-cycle sinusoidal functions, can be shown as

$$\Phi_{qocc}(f) = \frac{\cos^4 2\pi fT}{(1 - (4fT)^2)^4} \quad (19.36)$$

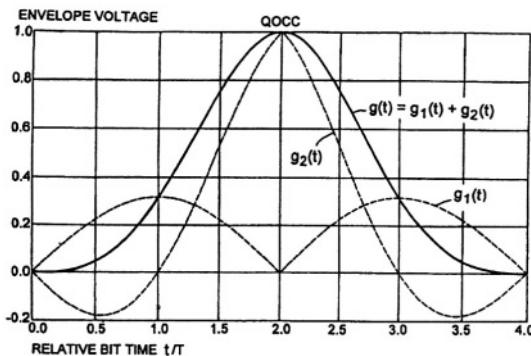


Figure 19.10. Pulse shaping waveform used in QOCC modulation.

from which it can be easily deduced that the PSD side lobe reduction rate for QOCC modulation is inversely proportional to $(fT)^{-8}$, which is much faster than that for MSK, $(fT)^{-4}$. The PSD's for both QOCC and MSK are compared in Figure (19.11).

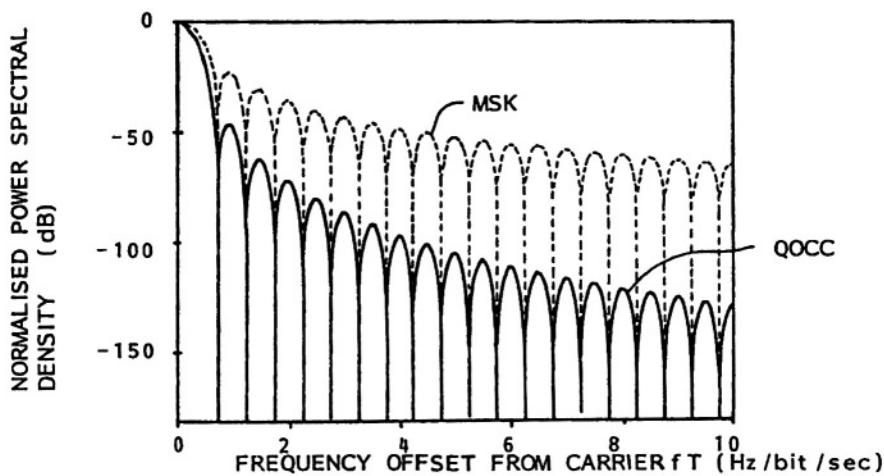


Figure 19.11. PSD function of QOCC in comparison with that of MSK.

Now, let us illustrate another example in using time-domain convolution method to design PSW. This time, we would like to start with two different $2T$ -long waveforms, one being again half-cycle sinusoidal function $g_{mask}(t)$ defined in (19.34) and the other being $2T$ -duration QOTC

PSW given by

$$g_{qotc}^{\dagger}(t) = \begin{cases} \frac{t}{T} - \frac{1}{2\pi} \sin \frac{2\pi t}{T}, & 0 < t \leq T \\ 2 - \frac{t}{T} + \frac{1}{2\pi} \sin \frac{2\pi t}{T}, & T < t \leq 2T \end{cases} \quad (19.37)$$

Please note that (19.37) is slightly different from (19.23) because they are defined in different time durations.

The convolution between $g_{msk}(t)$ and $g_{qotc}^{\dagger}(t)$ results in a new normalized PSW, which is used in a new QO modulation scheme named as *Modified Quadrature-Overlapped Convolved Cosine* (MQOCC) modulation, as

$$g_{mqocc}(t) = \begin{cases} \frac{T \sin \frac{2\pi t}{T}}{15\pi^2} - \frac{64T \sin \frac{\pi t}{2T}}{15\pi^2} + \frac{2t}{\pi}, & 0 < t \leq T \\ -\frac{T \sin \frac{2\pi t}{T}}{15\pi^2} - \frac{128 \cos \frac{\pi t}{2T}}{15\pi^2} \\ -\frac{64T \sin \frac{\pi t}{2T}}{15\pi^2} + \frac{2(2T-t)}{\pi}, & T < t \leq 2T \\ \frac{T \sin \frac{2\pi t}{T}}{15\pi^2} - \frac{\sqrt{2T} \cos \pi(\frac{t}{2T} + \frac{1}{4})}{\pi} \\ -\frac{251T \cos \frac{\pi t}{2T}}{30\pi^2} + \frac{41T \sin \frac{6\pi t}{2T}}{10\pi^2} - \frac{4T-2t}{\pi}, & 2T < t \leq 3T \\ -\frac{T \sin \frac{2\pi t}{T}}{15\pi^2} + \frac{64T \sin \frac{\pi t}{2T}}{15\pi^2} + \frac{8T-2t}{\pi}, & 3T < t \leq 4T \end{cases} \quad (19.38)$$

whose normalized PSD function can be easily shown to be

$$\Phi_{mqocc}(f) = \left(\frac{\sin \pi f T}{\pi f T} \right)^4 \frac{\cos^2 2\pi f T}{(1 - (4fT)^2)^2 (1 - (fT)^2)^2} \quad (19.39)$$

The PSW and PSD for MQOCC modulation are shown in Figures (19.12) and (19.13) respectively.

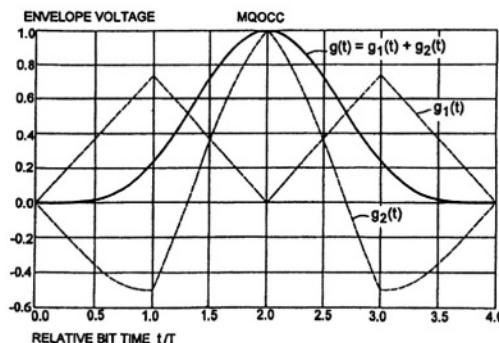


Figure 19.12. Pulse shaping waveform used in MQOCC modulation.

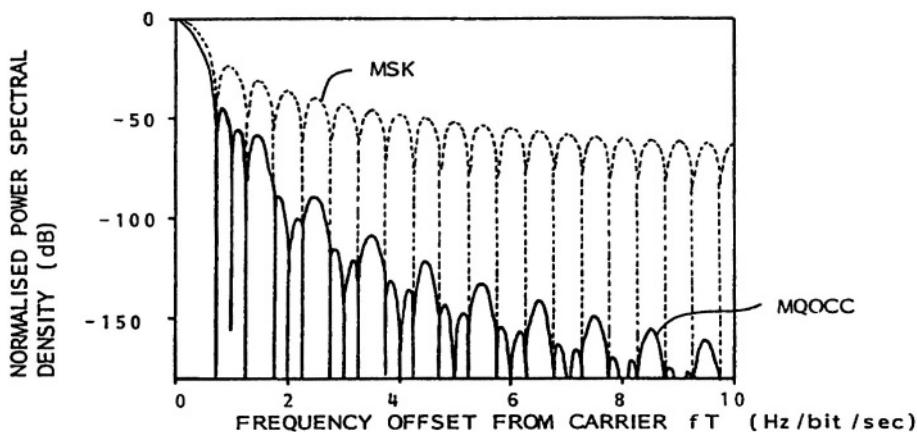


Figure 19.13. PSD function of MQOCC in comparison with that of MSK.

The TDC method retains several salient features in designing PSW's, which can be summarized as follows:

- It is extremely easy to ensure a specific requirement on PSD side lobes roll-off rate of obtained modulated signal. For instance, if you want to design a PSW with a particular side lobe roll-off rate as $(fT)^{-(m+n)}$, what you need to do is to find two suitable waveforms $f_1(t)$ and $f_2(t)$, each with its PSD side lobe roll-off rate as $(fT)^{-m}$ and $(fT)^{-n}$ respectively. Take the QOCC PSW as an example, which is obtained by convolution of two identical half-cycle sinusoidal waveforms or (19.34), each of which has its PSD side lobe roll-off rate as $(fT)^{-4}$, as seen from its PSD function

$$\Phi_{msk}(f) = \frac{\cos^2 2\pi f T}{(1 - (4fT)^2)^2} \quad (19.40)$$

Therefore, we will surely get the QOCC PSW with its PSD side lobe roll-off rate being $(fT)^{-8}$, which is the double of that of MSK. This feature in the TDC method gives us a great degree-of-freedom in controlling a suitable spectral efficiency of resultant PSW, otherwise hardly possible.

- The TDC method guarantees also a controllable PSD main lobe width in the target PSW, which is important when designing a bandwidth efficient digital modulation if some specific requirements on the main lobe exist. For example, if you have known the

PSD functions of two seed time-domain waveforms, $F_1(f)$ with its main lobe width being B_1 and $F_2(f)$ with its main lobe width being B_2 respectively, then you can conclude that the PSD main lobe of obtained PSW B must satisfy $B = \min(B_1, B_2)$. As many digital systems often take the main lobe of transmitted signal as the bandwidth of the bandpass filter, this feature in the TDC method can facilitate prediction of bandwidth requirement of a system that uses the obtained PSW.

- The procedure of the TDC method is very straight-forward and computationally simple. There is no complex computations involved, which can help in reducing computational complexity of the PSW design process. What you need to do is to collect as many seed waveforms as possible, which should be relatively simple in their representation and waveform appearance. A data base for those seed waveforms has been created by us, which include the following time functions, just list a few as examples (we assume all seed waveforms to be normalized and equal length of $2T$):

1 Square waveform:

$$g_{sqr}(t) = \begin{cases} 1, & 0 < t \leq 2T \\ 0, & \text{e. w.} \end{cases} \quad (19.41)$$

2 Triangular waveform:

$$g_{tri}(t) = \begin{cases} \frac{t}{T}, & 0 < t \leq T \\ 2 - \frac{t}{T}, & T < t \leq 2T \end{cases} \quad (19.42)$$

3 Half-cycle sinusoidal waveform:

$$g_{sin}(t) = \begin{cases} \sin \frac{\pi t}{2T}, & 0 < t \leq 2T \\ 0, & \text{e. w.} \end{cases} \quad (19.43)$$

4 Reverse half sinusoidal waveform:

$$g_{rhs}(t) = \begin{cases} 1 - \sin \frac{\pi t}{2T}, & 0 < t \leq 2T \\ 0, & \text{e. w.} \end{cases} \quad (19.44)$$

5 Raised-cosine waveform:

$$g_{rc}(t) = \begin{cases} \frac{1}{2}(1 - \cos \frac{\pi t}{T}), & 0 < t \leq 2T \\ 0, & \text{e. w.} \end{cases} \quad (19.45)$$

6 Reverse raised-cosine waveform:

$$g_{rrc}(t) = \begin{cases} \frac{1}{2}(1 + \cos \frac{\pi t}{T}), & 0 < t \leq 2T \\ 0, & \text{e. w.} \end{cases} \quad (19.46)$$

7 Back-to-back sinusoidal waveform:

$$g_{bbs}(t) = \begin{cases} -\sin \frac{2\pi t}{T}, & 0 < t \leq T \\ \sin \frac{2\pi t}{T}, & T < t \leq 2T \end{cases} \quad (19.47)$$

8 Dual half sinusoidal waveform:

$$g_{dhs}(t) = \begin{cases} \sin \frac{\pi t}{T}, & 0 < t \leq T \\ -\sin \frac{\pi t}{T}, & T < t \leq 2T \end{cases} \quad (19.48)$$

4. Merit Parameters of Pulse-Shaped Modems

To objectively measure the performance of a pulse-shaped modulation, we need to define some suitable merit parameters.

Since we have already introduced four new QO modulations, QOTC, MQOTC, QOCC and MQOCC, which were obtained by using various pulse shaping techniques, it is of great interest to us to make a comparison among them with some other conventional digital modulation schemes, such as QPSK and MSK, etc.

Let us first define *out-of-band power* (OOBP) to measure the bandwidth efficiency of different modulations as

$$P_{out}(B) = 10 \log \left[1 - \frac{\int_{-B}^B \Phi(f) df}{\int_{-\infty}^{\infty} \Phi(f) df} \right] \quad (dB) \quad (19.49)$$

where B is the single-sided bandwidth of the modulated signal of interest and $\Phi(f)$ is its PSD function. Obviously, the value in $P_{out}(B)$ tells us how much power emission outside a particular bandwidth ($-B, B$), which should be made as small as possible for bandwidth efficiency enhancement. Figure (19.14) shows the OOBP's for all four new QO modulation schemes proposed by us, which are compared to that of MSK.

The bandwidth efficiency of different modulations can also be compared using their normalized single-sided bandwidths, as shown in Table (19.2).

Please note that all data shown in Table (19.2) were obtained with the duration of the PSW of concern defined within $(0, 4T)$. The results show a significant improvement in single-sided bandwidth (fT) if the proposed modulations can be used to replace the legacy schemes, such as OQPSK and MSK.

The eye-pattern for QOTC modulation is shown in Figure (19.15), where it is clearly shown that there is a flat opening width in $(-0.1t/T, 0.1t/T)$

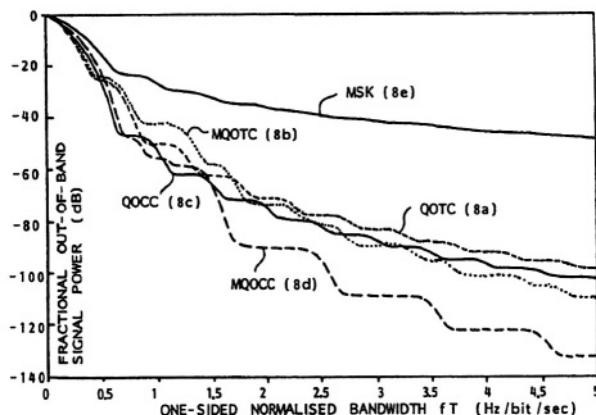


Figure 19.14. OOBP of QOTC, MQOTC, QOCC and MQOCC modulations in comparison with that of MSK.

Table 19.2. Comparison of normalized single-sided bandwidths (in fT) for different digital modulations with varying normalized PSD levels.

| PSD (dB) | QPSK | MSK | QORC | QOTC | MQOTC | QOCC | MQOCC |
|----------|---------|---------|--------|--------|--------|--------|--------|
| -3 | 0.2211 | 0.2968 | 0.1798 | 0.1904 | 0.1933 | 0.2124 | 0.2365 |
| -10 | 0.369 | 0.5095 | 0.314 | 0.3301 | 0.3341 | 0.3768 | 0.4192 |
| -20 | 1.3406 | 0.66517 | 0.4124 | 0.4274 | 0.4305 | 0.5097 | 0.5648 |
| -30 | 4.7987 | 1.1419 | 0.3989 | 0.7664 | 0.7901 | 0.5948 | 0.6538 |
| -40 | 15.7714 | 2.51 | 0.6989 | 0.8623 | 0.8785 | 0.6517 | 0.7064 |
| -50 | 49.283 | 4.4887 | 1.065 | 1.2426 | 1.4024 | 0.8517 | 0.9245 |
| -60 | 158.739 | 7.5661 | 1.6609 | 1.3763 | 1.712 | 1.1419 | 1.5055 |
| -70 | 503.252 | 14.018 | 2.4137 | 1.8653 | 1.8168 | 1.6200 | 1.6308 |
| -80 | 1591.25 | 25.001 | 3.6449 | 2.7927 | 2.4246 | 2.5100 | 1.693 |
| -90 | 5032.75 | 44.031 | 5.3857 | 3.7673 | 2.8084 | 3.0876 | 2.4943 |
| -100 | 15915.8 | 78.991 | 7.8884 | 4.816 | 3.7687 | 4.0903 | 2.6503 |

of the eye-pattern that is important for QOTC to have a robust performance against sampling timing jitter.

The phaser diagram of QOCC modulation is shown in Figure (19.16), which is to show how much fluctuation in its modulated signal envelope. For a constant-envelope signal, the phaser diagram should be a perfect circle. To characterize the fluctuation in modulated signal's envelope, we have introduced a parameter $0 \leq \rho \leq 1$, named as *envelope fluctuation factor*, whose value should be made as close to unit as possible to reduce the envelope fluctuation in modulated signal. Clearly, this parameter has

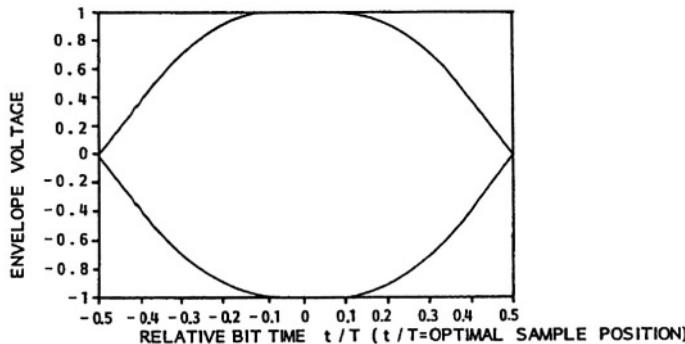


Figure 19.15. Eye-pattern of QOTC modulations.

a great impact on how a modem will perform in a non-linear channels, such as satellite channels and mobile cellular systems, where non-linear C-class power amplifiers are widely used. Such power amplifiers work in a very similar way as a hard-limiter, such that all information carried in the envelope or amplitude will be lost, causing a serious distortion to the modulated signal.

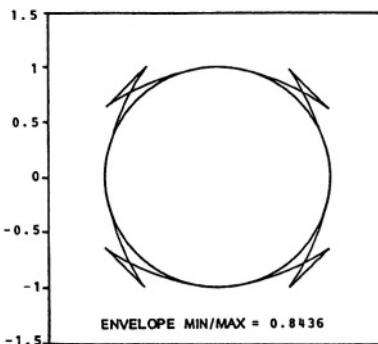


Figure 19.16. Phaser diagram of QOCC modulation.

Another important parameter relevant to pulse-shaping waveforms is defined as

$$\mu = \frac{1}{4T} \int_0^{4T} g^2(t) dt \quad (19.50)$$

where it has been assumed that the normalized PSW, $g(t)$, of concern has a duration of $4T$. In fact, (19.50) gives the average power of a PSW, which is also important in determining the shapes of a pulse waveform. In some cases we would like also to know the sampling value at a quarter of the pulse duration ($4T$), or $g(T)$, which decides the opening of the eye-pattern of a PSW. Thus, we have three important parameters in our hand: ρ , μ and $g(T)$, all of which will be used to characterize the exact shape of a particular PSW. Table (19.3) gives their values for four different QO modulation schemes using distinct PSW's.

Table 19.3. Three shape parameters (ρ , μ and $g(T)$) for QORC, QOTC, MQOTC, QOCC and MQOCC modulations

| Modulation-type | ρ | μ | $g(T)$ |
|-----------------|--------|--------|--------|
| QORC | 0.7071 | 0.375 | 0.5 |
| QOTC | 0.7071 | 0.3967 | 0.5 |
| MQOTC | 0.7071 | 0.4024 | 0.5 |
| QOCC | 0.8436 | 0.2933 | 0.3183 |
| MQOCC | 0.9041 | 0.2693 | 0.2363 |

It is seen from Table (19.3) that MQOCC has its *envelope fluctuation factor* reaching $\rho = 0.9041$, which is very close to unit to become a constant envelope modulation scheme. Therefore, it should be suitable for the applications in non-linear channels, including mobile cellular systems where all handset transmitters always use C-class RF amplifiers with high power-efficiency due to the concern of battery stand-by time.

5. Spectral Efficiency in Band-Limited Nonlinear Channels

The bandwidth-efficiency of a digital modem is always an important issue whenever a practical application is concerned. In this section we will offer a comprehensive discussion on how to evaluate spectral efficiency of a pulse-shaped QO modulation over band-limited nonlinear channels, which can be found in many applications such as satellite communications and mobile cellular systems.

5.1 Four More QO Modulations

It should be pointed out that the evaluation of spectral efficiency for a digital modem over a band-limited nonlinear channels can not be done

using only the PSD function of the modulated signal. Instead, the study should take band-limited nonlinear channels into account and it can be a complex procedure. In the text followed there are four QO modulations involved (whose modulator structure retains the same as that shown in Figure (19.1)), which are

- 1 *overlapped minimum shift keying* (OMSK) modulation, whose PSW can be written as

$$g_{omsk}(t) = \begin{cases} \frac{1}{\pi} \sin \frac{\pi t}{2T} - \frac{t}{2T} \cos \frac{\pi t}{2T}, & 0 < t \leq 2T \\ (\frac{t}{2T} - 2) \cos \frac{\pi t}{2T} - \frac{1}{\pi} \sin \frac{\pi t}{2T}, & 2T < t \leq 4T \end{cases} \quad (19.51)$$

- 2 *minimum shift keying triangular cosine* (MSKTC) modulation with its PSW being

$$g_{msktc}(t) = \begin{cases} \frac{T \sin \frac{2\pi t}{T}}{15\pi^2} - \frac{64T \sin \frac{\pi t}{2T}}{15\pi^2} + \frac{2t}{\pi}, & 0 < t \leq T \\ -\frac{T \sin \frac{2\pi t}{T}}{15\pi^2} - \frac{128T \cos \frac{\pi t}{2T}}{15\pi^2} \\ -\frac{64T \sin \frac{\pi t}{2T}}{15\pi^2} + \frac{2(2T-t)}{\pi}, & T < t \leq 2T \\ \frac{T \sin \frac{2\pi t}{T}}{15\pi^2} - \frac{\sqrt{2}T \cos[\pi(\frac{t}{2T} + \frac{1}{4})]}{6\pi^2}, & 2T < t \leq 3T \\ -\frac{251T \cos \frac{\pi t}{2T}}{30\pi^2} + \frac{41T \sin \frac{\pi t}{2T}}{10\pi^2} - \frac{2(2T-t)}{\pi}, & 3T < t \leq 4T \\ -\frac{T \sin \frac{2\pi t}{T}}{15\pi^2} + \frac{64T \sin \frac{\pi t}{2T}}{15\pi^2} + \frac{2(4T-t)}{\pi}, & 3T < t \leq 4T \end{cases} \quad (19.52)$$

which should be normalized by a factor of $\frac{128T}{15\pi^2}$;

- 3 *raised-cosine triangular cosine* (RCTC) modulation, whose PSW is given by

$$g_{rctc}(t) = \begin{cases} -T \frac{\cos \frac{2\pi t}{T}}{24\pi^2} + 2T \frac{\cos \frac{\pi t}{T}}{3\pi^2} \\ -\frac{5T}{8\pi^2} + \frac{t^2}{4T}, & 0 < t \leq T \\ T \frac{\cos \frac{2\pi t}{T}}{24\pi^2} + 2T \frac{\cos \frac{\pi t}{T}}{\pi^2} \\ + \frac{5T}{8\pi^2} - \frac{T}{2} - \frac{t^2}{4T} + t, & T < t \leq 3T \\ -t \frac{\cos \frac{2\pi t}{T}}{24\pi^2} + 2T \frac{\cos \frac{\pi t}{T}}{3\pi^2} \\ -\frac{5T}{8\pi^2} + 4T + \frac{t^2}{4T} - 2t, & 3T < t \leq 4T \end{cases} \quad (19.53)$$

which can be normalized by a factor of $\frac{8T}{3\pi^2} + \frac{T}{2}$; and

4 sinusoidal quadrature overlapped triangular cosine (SQOTC) modulation, whose PSW can be represented as

$$g_{sqotc}(t) = \begin{cases} \frac{t^3}{6T^2} - \frac{t}{2\pi^2} + \frac{5T}{2(2\pi)^3} \sin \frac{2\pi t}{T} \\ \quad - \frac{t}{2(2\pi)^3} \cos \frac{2\pi t}{T}, & 0 < t \leq T \\ \frac{2T}{3} - 2t + \frac{2t^2}{T} - \frac{t^3}{2T^2} + \frac{6t-8T}{(2\pi)^2} \\ - \frac{15T}{2(2\pi)^3} \sin \frac{2\pi t}{T} + \frac{(\frac{3t}{2}-2T)}{(2\pi)^2} \cos \frac{2\pi t}{T}, & T < t \leq 2T \\ - \frac{22T}{3} + 10t - \frac{4t^2}{T} + \frac{t^3}{2T^2} + \frac{16T-6t}{(2\pi)^2} \\ + \frac{15T}{2(2\pi)^3} \sin \frac{2\pi t}{T} + \frac{(4T-\frac{3t}{2})}{(2\pi)^2} \cos \frac{2\pi t}{T}, & 2T < t \leq 3T \\ \frac{32T}{3} - 8t + \frac{2t^2}{T} - \frac{t^3}{6T} + \frac{2t-8T}{(2\pi)^2} \\ - \frac{5T}{2(2\pi)^3} \sin \frac{2\pi t}{T} + \frac{(\frac{t}{2}-2T)}{(2\pi)^2} \cos \frac{2\pi t}{T}, & 3T < t \leq 4T \end{cases} \quad (19.54)$$

whose normalizing factor is $\frac{2T}{3} + \frac{5T}{4\pi^2}$.

All above four PSW's can be generated with the help of the TDC method by choosing some proper seed pulse waveforms, as explained in the previous section. The four different PSW's are shown in Figure (19.17), where the raised-cosine waveform is also plotted for easy comparison.

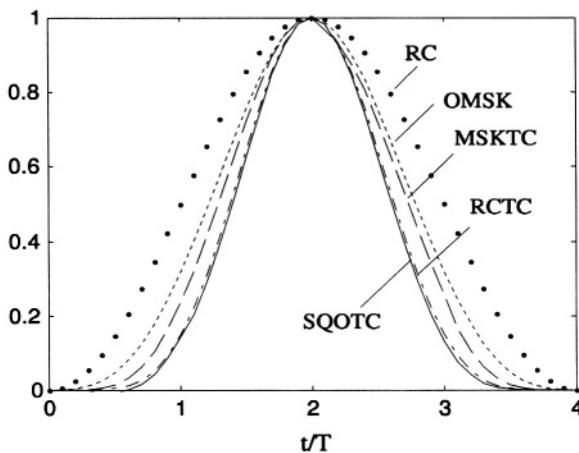


Figure 19.17. Pulse-shaping waveforms for OMSK, MSKTC, RCTC and SQOTC modulations, which are compared with that of QORC.

5.2 Band-Limited Nonlinear Channel Model

The QO modulation system model considering the non-linear channel is illustrated in Figure (19.18), where a binary input data stream $\{c_k\}$, with bit duration T , is demultiplexed by a serial-to-parallel converter into two independent binary data sequences $\{a_k\}$ and $\{b_k\}$. The data sequences $\{a_k\}$ and $\{b_k\}$ (each in non-return-to-zero (NRZ) format taking the values of ± 1) then pass through pulse-shaping filters, $p(t)$ and $q(t)$, and are modulated onto the quadrature carriers, resulting in the following transmitted signal

$$\begin{aligned} s(t) = & \left[\sum_{n=-\infty}^{\infty} a_n p(t - 2nT) \right] \cos \omega_c t \\ & + \left[\sum_{n=-\infty}^{\infty} b_n q(t - 2nT) \right] \sin \omega_c t \quad (a_n, b_n = \pm 1) \end{aligned} \quad (19.55)$$

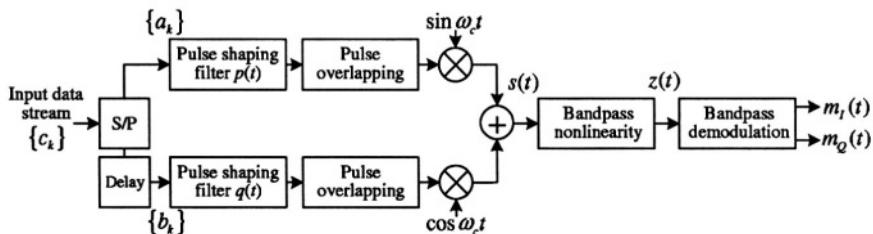


Figure 19.18. A generic system model for studying spectral efficiency of any quadrature modulation schemes in band-limited nonlinear channels, valid for overlapped or non-overlapped (without two *pulse overlapping* elements) and staggered or non-staggered (without the *delay* element) modulations.

The spectral characteristics of the modulation are primarily determined by the pulse-shaping filters $p(t)$ and $q(t)$ in I and Q channels, respectively. Pulse-shaping can be implemented in two ways; overlapped and non-overlapped, each of which can be either staggered or non-staggered, depending on the relative phase shift between I and Q channel bit streams. The non-overlapped modulations include QPSK, offset QPSK and MSK modulations. As for the overlapped modulations, two consecutive bit pulses in both I and Q channels are overlapped by just half of the pulse duration ($2T$), as in the QORC and QOTC modulations. In staggered modulation schemes, the Q channel bit stream is delayed by T with respect to that in the I channel. In fact, all four new modulations concerned in this section are the staggered (offset) QO modulations and the duration of a symbol pulse in either I or Q channels is $4T$, where T is the bit duration of the input data stream $\{c_k\}$.

Although the pulse-shape functions of the four new QO modulations, given in (19.51) to (19.54), look quite complicated, they can be realized using digital filter technology. Because the pulse shaping waveforms can be viewed as the output from a filter, whose input is the binary data waveform (normally, antipodal square pulses taking the levels of -1 or +1), the impulse response of the pulse-shaping filter can be easily obtained with the help of digital filter design theory. Two such pulse-shaping filters are needed in a QO modem, one for the I channel and another for the Q channel. Due to the overlapping of consecutive bit symbols in a QO modem, a waveform buffer, together with a delay and addition unit, is required for both I and Q channels. This means that the output pulse-shaped signals can be appropriately overlapped according to the signs of the corresponding consecutive bits. The complexity of such a QO modem is slightly greater than that of a non-overlapped modem, such as MSK. However, there is no technical difficulties in implementing such a QO modem and the hardware complexity can be kept reasonably low. We have to admit that the hardware implementation of the proposed modems has some impact on the real-world spectral performance of the modems, due to certain realization problems, such as quantization errors existing in realizing digital filters, etc. Therefore, the performance analysis in this section only reflects the results of an ideal scenario without considering all those realization issues. On the other hand, as the same ideal scenario is applied to all compared modulation schemes, the relative difference in their performance also tells which one is better.

As $s(t)$ is to be transmitted over a non-linear channel, it is convenient for us to represent $s(t)$ in (19.55) in the equivalent polar form as

$$s(t) = \sum_{k=-\infty}^{\infty} R_T(t - 2kT) \cos(\omega_c t - \eta_T(t - 2kT)) \quad (19.56)$$

where $R_T(t)$ and $\eta_T(t)$ have to be determined and are non-zero in the interval $[0,2T]$. Then, $R_T(t - 2kT)$ and $\eta_T(t - 2kT)$ are just the time-shifted versions of $R_T(t)$ and $\eta_T(t)$, respectively. For the staggered and overlapped modulations, the expressions for $R_T(t - 2kT)$ and $\eta_T(t - 2kT)$ are given as follows (its proof has been given in [39, 41]):

$$\begin{aligned} & R_T(t - 2kT) \\ &= \begin{cases} \left\{ \left[\sum_{i=0}^1 a_{k-i} p(t - 2(k-i)T) \right]^2 \right. \\ \left. + \left[\sum_{i=0}^2 b_{k-i} q(t - 2(k-i)T) \right]^2 \right\}^{1/2}, & 2kT \leq i \leq 2(k+1)T \\ 0, & \text{e.w.} \end{cases} \end{aligned} \quad (19.57)$$

and

$$\eta_T(t - 2kT) = \begin{cases} \tan^{-1} \left[\frac{\sum_{i=0}^2 b_{k-i}(q(t-2(k-i)T))}{\sum_{i=0}^1 a_{k-i}(p(t-2(k-i)T))} \right], & 2kT \leq i \leq 2(k+1)T \\ 0, & \text{e.w.} \end{cases} \quad (19.58)$$

The non-linearity has equivalent AM-AM and AM-PM characteristics denoted by $\tilde{f}(\cdot)$ and $\tilde{g}(\cdot)$ [39, 41], respectively. The output signal from the non-linear channel can be written as

$$\begin{aligned} z(t) &= \sum_{k=-\infty}^{\infty} \tilde{f}(R_T(t - 2kT)) \cos [\omega_c t - \eta_T(t - 2kT) - \tilde{g}(R_T(t - 2kT))] \\ &= m_I(t) \cos \omega_c t + m_Q(t) \sin \omega_c t \end{aligned} \quad (19.59)$$

where the in-phase component $m_I(t)$ can be written into

$$m_I(t) = \sum_{k=-\infty}^{\infty} d_I(t - 2kT; a_k(1), b_k(2)) \quad (19.60)$$

with

$$\begin{aligned} d_I(t - 2kT; a_k(1), b_k(2)) \\ = \begin{cases} \tilde{f}(R_T(t - 2kT)) \cos[\eta_T(t - 2kT) + \tilde{g}(R_T(t - 2kT))], & 2kT \leq t \leq 2(k+1)T \\ 0, & \text{otherwise} \end{cases} \end{aligned} \quad (19.61)$$

and

$$a_k(1) = [a_{k-1}, a_k] \quad (19.62)$$

$$b_k(2) = [b_{k-2}, b_{k-1}, b_k] \quad (19.63)$$

Similarly, the quadrature demodulated signal $m_Q(t)$ over the non-linear channel can also be derived. Based on these equations, we can evaluate the PSD of the QO modulated signal over a band-limited nonlinear channel in the sequel.

5.3 PSD Expression Derivation

In this subsection, we shall first derive a generic PSD expression for the signals over a non-linear channel. From the definition, the power spectral density of $m_I(t)$ is given by

$$S_m(f) = \mathcal{F}(R_m(\tau)) = \mathcal{F}(\langle \overline{m_I(t)m_I(t-\tau)} \rangle) \quad (19.64)$$

where $R_m(\tau)$ is the autocorrelation function of the signal $m_I(t)$, \mathcal{F} denotes the Fourier transform operator, the over-bar denotes the statistical average and the symbol " $\langle \rangle$ " denotes the time average. Substituting equation (19.60) into (19.64), we have

$$S_m(f) = \overline{\mathcal{F}\left\{\left[\sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \overline{d(t-2jT; a_j(1), b_j(2))d(t-2kT-\tau; a_k(1), b_k(2))}\right]\right\}} \quad (19.65)$$

Denoting

$$v(t-2jT) = d(t-2jT; a_j(1), b_j(2)) \quad (19.66)$$

$$v(t-2kT-\tau) = d(t-2kT-\tau; a_k(1), b_k(2)) \quad (19.67)$$

we thus obtain the following equation to calculate the time average for (19.65)

$$\begin{aligned} & \left\langle \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} v(t-2jT)v(t-2kT-\tau) \right\rangle \\ &= \langle \cdots + v(t)v(t-\tau) + v(t)v(t-2T-\tau) + \cdots \\ &+ v(t-2T)v(t-\tau) + v(t-2T)v(t-2T-\tau) + \cdots \rangle \\ &= \frac{1}{2T} \sum_{k=-\infty}^{\infty} \int_0^{2T} v(t)v(-(t-\tau)-2kT)dt \\ &= \frac{1}{2T} \sum_{k=-\infty}^{\infty} v(t) \otimes v(-\tau-2kT) \end{aligned} \quad (19.68)$$

where \otimes denotes the time convolution operation. Thus, the power spectral density in (19.65) can be reduced to

$$\begin{aligned} S_m(f) &= \frac{1}{2T} \mathcal{F}\left\{\sum_{k=-\infty}^{\infty} \overline{d(\tau; a_0(1), b_0(2)) \otimes d(-\tau-2kT; a_k(1), b_k(2))}\right\} \\ &= \frac{1}{2T} \sum_{k=-\infty}^{\infty} \overline{D(f; a_0(1), b_0(2)) D^*(f; a_k(1), b_k(2))} e^{j4\pi f k T} \end{aligned} \quad (19.69)$$

where the asterisk denotes the complex conjugate, and

$$\begin{aligned} D(f; a_k(1), b_k(2)) &= \int_0^{\infty} d(t; a_k(1), b_k(2)) e^{-j2\pi ft} dt \\ &= \int_0^{2T} d(t; a_k(1), b_k(2)) e^{-j2\pi ft} dt \end{aligned} \quad (19.70)$$

We should bear in mind that the PSD expression given in (19.69) is valid for staggered and overlapped signals passing through a non-linear channel. When the non-linearity is modeled by a hard limiter, the AM-AM and AM-PM characteristics of (19.59) become

$$\tilde{f}(\cdot) = \gamma = \sqrt{2} \quad (19.71)$$

$$\tilde{g}(\cdot) = 0 \quad (19.72)$$

Applying Equation (19.72) to (19.61), we obtain a further simplified expression

$$\begin{aligned} d_I(t - 2kT; a_k(1), b_k(2)) \\ = \begin{cases} \gamma \cos[\eta_T(t - 2kT)], & 2kT \leq t \leq 2(k+1)T \\ 0, & \text{otherwise} \end{cases} \end{aligned} \quad (19.73)$$

In the above analysis, we have only considered the I channel signal $m_I(t)$. In fact, the same analysis can also be applied to the Q channel signal, $m_Q(t)$. Therefore, we could ignore the subscripts I and Q and (19.69) gives a generic PSD expression for the QO signals.

5.4 Numerical Evaluation of PSD

Equation (19.69), however, is not suitable for a direct PSD numerical calculation for a specific QO signal. Therefore, we want to derive a computationally efficient PSD expression suitable for any staggered or/and overlapped modulations. Using the results obtained in [39, 41], the numerical expression for the PSD of staggered and overlapped modulations (assuming that pulse-shaping waveform $p(t)$ is symmetrical about $t = 2T$) after passing through a non-linear channel is given below

$$\begin{aligned} S_m(f) = & 2T\gamma^2 \left\{ \frac{3}{4} \sum_{k=1}^4 R_k^2 + \frac{1}{4} \sum_{k=1}^4 I_k^2 \right. \\ & - \frac{1}{2} (R_1 R_3 + R_2 R_4 - I_1 I_3 - I_2 I_4) \\ & + \frac{1}{8} \cos 8\pi f T \left[\left(\sum_{k=1}^4 R_k \right)^2 - \left(\sum_{k=1}^4 I_k \right)^2 \right] \\ & - \frac{1}{4} \sin 8\pi f T \left[\left(\sum_{k=1}^4 R_k \right) \left(\sum_{k=1}^4 I_k \right) \right] \\ & \left. + \frac{1}{2} (R_1 + R_2 - R_3 - R_4) \right\} \end{aligned}$$

$$\times \left(\sum_{k=1}^4 R_k \cos 4\pi fT - \sum_{k=1}^4 I_k \sin 4\pi fT \right) \quad (19.74)$$

where

$$\begin{aligned} R_1 &= Re[K(f; 1, 1)]; & I_1 &= Im[K(f; 1, 1)] \\ R_2 &= Re[K(f; 1, -1)]; & I_2 &= Im[K(f; 1, -1)] \\ R_3 &= -Re[K(f; -1, 1)]; & I_3 &= -Im[K(f; -1, 1)] \\ R_4 &= -Re[K(f; -1, -1)]; & I_4 &= -Im[K(f; -1, -1)] \end{aligned} \quad (19.75)$$

and

$$\begin{aligned} K(f; x_k, y_k) \\ = \frac{1}{2T} \int_0^T \frac{[p(t + 2T) + x_k p(t)] e^{-j2\pi f t} dt}{\{[p(t + 2T) + x_k p(t)]^2 + [q(t + 2T) + y_k q(t + 4T)]^2\}^{1/2}} \end{aligned} \quad (19.76)$$

We have also defined x_k and y_k as

$$\begin{aligned} y_k &= b_k b_{k-1} = b_k / b_{k-1} = b_{k-1} / b_k \\ x_k &= a_k a_{k-1} = a_k / a_{k-1} = a_{k-1} / a_k \end{aligned} \quad (19.77)$$

where $\{a_k\}$ and $\{b_k\}$ take the values of ± 1 with equal likelihood. Then x_k and y_k are identically independently distributed (*i.i.d.*) sequences, taking values of ± 1 with equal probability. In the next subsection, (19.74) will be used to study the spectral efficiency of the four new QO modulations over the non-linear channels.

5.5 Numerical Results

Equation (19.74) can be used to numerically evaluate the spectral efficiency for various modulations (either overlapped or non-overlapped, staggered or non-staggered). Here we are particularly interested in the staggered QO modulations over both linear and non-linear channels.

5.5.1 PSD over Linear Channels. The PSD's of the QO signals transmitted over a linear channel are obtained by performing Fourier transform of the autocorrelation functions of transmitted signals. The pulse-shaping waveforms of staggered QORC, OMSK, MSKTC, RCTC and SQOTC modulations have been shown in Figure (19.17); whereas their corresponding spectra are given in Figure (19.19).

It can be observed in Figure (19.19) that the new modulation schemes over the linear channel have faster side lobe reduction rates than the

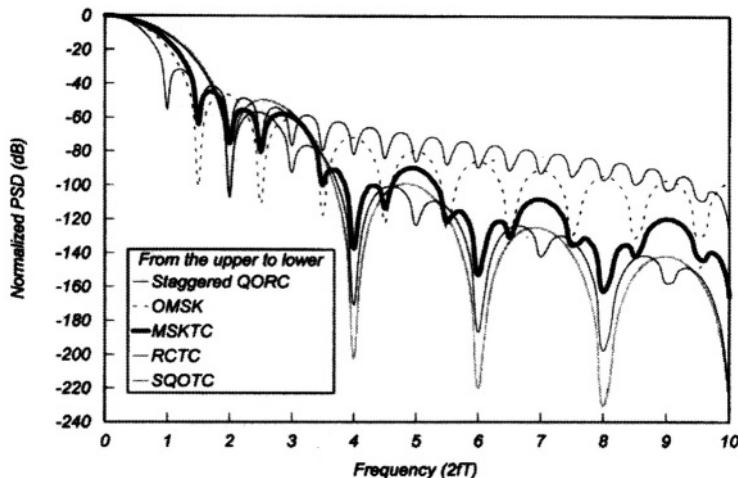


Figure 19.19. PSD functions of staggered QORC, OMSK, MSKTC, RCTC and SQOTC modulated signals in a linear channel without bandpass filter.

staggered QORC modulation, which already possesses better spectral characteristics than the OQPSK and MSK modulations [9, 17]. The PSD main lobe widths of the OMSK and MSKTC signals remain the same as that of MSK, which is 1.5 times wider than that of staggered QORC modulation. As for the RCTC and SQOTC signals, their PSD main lobe widths are twice as wide as the staggered QORC modulation. However, they exhibit faster side lobe reduction rate than the QORC modulation.

It is interesting to note that somehow there is a relation between the PSD and the shaping-pulse appearance. The flatter the central part of the bit pulse, the narrower the PSD main lobe; whereas the flatter the part around the edges, the faster the side lobe levels drop. It is verified in Figures (19.17) and (19.19) that the RCTC and SQOTC pulses with flatter edges have faster side lobe reduction rates, and the staggered QORC pulse has a narrower main lobe due to its flat central portion of the pulse shape.

5.5.2 PSD over Nonlinear Channels.

Using Equation (19.74), we can obtain the PSD's of staggered and overlapped modulations (staggered QORC, OMSK, MSKTC, RCTC and SQOTC modulations) after passing through the hard-limited channel. The power spectra of the OMSK, MSKTC, RCTC and SQOTC modulations, prior

to and after hard-limiting, are shown in Figures (19.20) and (19.21) respectively. After hard-limiting, the spectral side lobes of the OMSK and MSKTC modulations can maintain their respective levels prior to hard-limiting. However, the side lobe levels regenerated due to the hard-limiting effect for the RCTC and SQOTC modulations are higher than those of the OMSK and MSKTC modulations, although they are still better than the QORC modulation.

The power spectra after bandpass and hard-limiting channel of the OQPSK, MSK, staggered QORC and the four new modulations are compared in Figures (19.22) and (19.23), respectively. As expected, the spectra of constant-envelope OQPSK and MSK modulations change little after hard-limiting (their PSDs are derived in [39, 41]). These modulations suffer from very high side lobe levels even prior to hard-limiting. On the other hand, the new QO modulations can keep relatively low side lobe levels (by taking into account the side lobe regeneration effect), if compared with those for OQPSK, MSK and staggered QORC modulations. From Figures (19.22) and (19.23), we can see that the MSKTC modulation yields the lowest side lobe level after bandpass hard-limiting channel. Figure (19.19) shows that RCTC and SQOTC modulations have the lowest side lobe levels before hard-limiting, because the PSD side lobes of the SQOTC and RCTC modulations are regenerated due to the hard-limiting effect. In general, the new modulations can offer a better spectral efficiency than that of their counter-part modulations OQPSK, MSK and staggered QORC.

6. Robustness Against Sampling Timing Jitter

It is well-known that QORC modulation [17] is more bandwidth-efficient than QPSK and MSK due mainly to the application of its pulse shaping and overlapping before quadrature carrier modulation. With the same transmission rate, the pulse duration in a QO modulation can be made twice as wide as that for QPSK or MSK [9, 15, 17]. In a QO modulation there should be no inter-symbol interference if sampling to received signal can be perfectly synchronized at $t = 2nT$ in both I and Q channels at a receiver, where n is any integer, T is the bit duration of the input signal to the modulator and raised-cosine pulse is defined over $(0, 4T)$ as given in (19.11). However, when timing for sampling is jittered because of either non-perfect bit synchronization or carrier phase tracking problems or other interference, the sampling from adjacent bits will degrade the bit error rate (BER) performance. Therefore, the shape

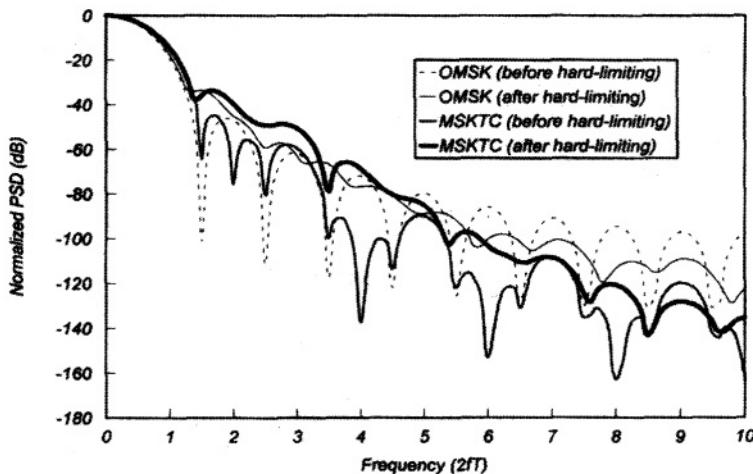


Figure 19.20. PSD functions of staggered OMSK and MSKTC modulated signals before and after bandpass hard-limiting channel.

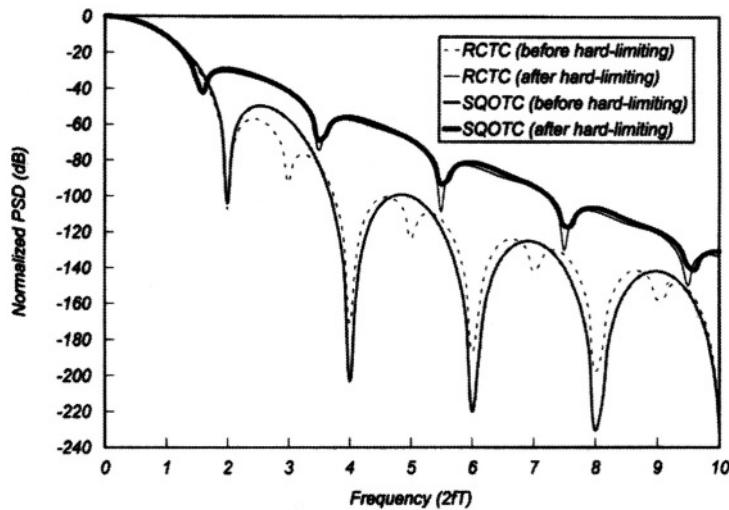


Figure 19.21. PSD functions of staggered RCTC and SQOTC modulated signals before and after bandpass hard-limiting channel.

of the pulse shaping waveform in a quadrature-overlapped modulation plays an important role in BER when timing jitter is present.

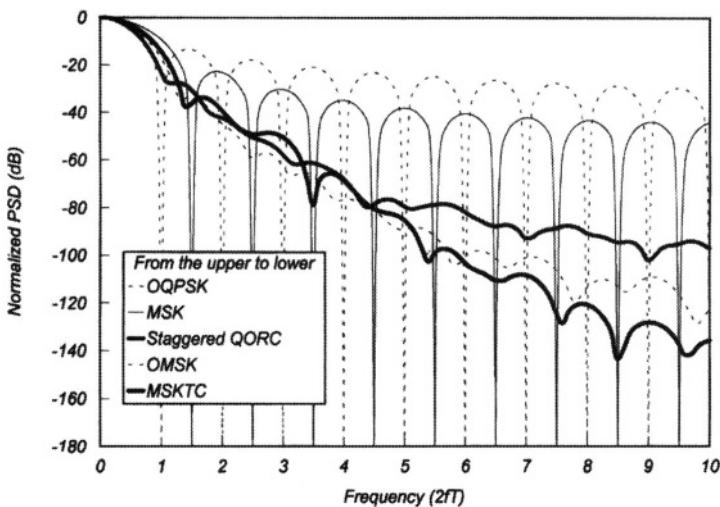


Figure 19.22. PSD functions of staggered OMSK and MSKTC modulated signals, which are compared to those for OQPSK and MSK modulated signals, after bandpass hard-limiting channel.

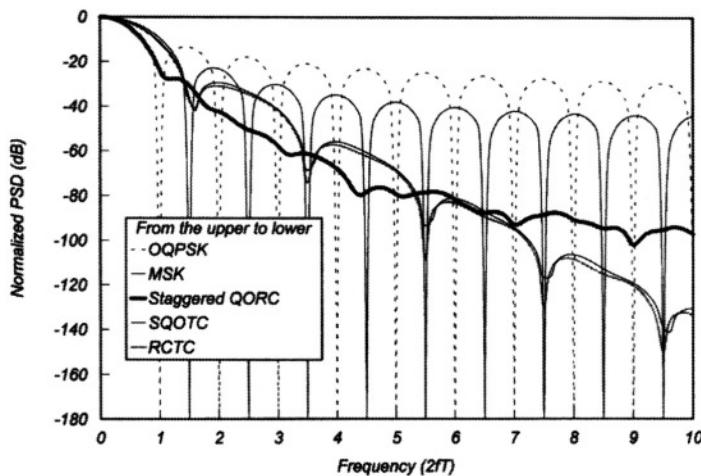


Figure 19.23. PSD functions of staggered RCTC and SQOTC modulated signals, which are compared to those for OQPSK and MSK modulated signals, after bandpass hard-limiting channel.

In this section we would like to study the impact of receiver timing jitter on the bit error rate (BER) of a pulse-shaped QO modulation.

We are interested in *Raised-Cosine and Triangular-Cosine* (RCTC) modulation, whose PSW has been given in (19.53). In fact, the RCTC waveform in (19.53) is generated with the help of the TDC method by time convolution between a raised-cosine (19.45) and a triangular-cosine (19.37) waveforms, both having a $2T$ -length. The resultant RCTC PSW, together with that of raised-cosine waveform, has been shown in Figure (19.17).

It can be seen from Figure (19.17) that the edges of the RCTC pulse (the regions nearby either $t=0$ or $4T$) are flatter than those of the raised-cosine pulse, thus a lower sampling interference level is expected when timing jitter exists. To concentrate on BER for QORC and RCTC modulations with timing jitter, inter-symbol interference (ISI) due to limited bandwidth is not considered here. The transmitted signal can be written as

$$s(t) = I(t) \cos \omega t + Q(t) \sin \omega t \quad (19.78)$$

where

$$I(t) = A \sum_{n=-\infty}^{\infty} a_n g(t - 2nT) \quad (19.79)$$

$$Q(t) = A \sum_{n=-\infty}^{\infty} b_n g(t - 2nT - T) \quad (19.80)$$

are baseband signals in the I and Q channels, respectively, a_n and b_n take either 1 or -1, A is the signal amplitude and $g(t)$ is the pulse shaping waveform defined within $(0, 4T)$. After corruption by Gaussian noise $n(t) = n_c \cos \omega t - n_s \sin \omega t$ with zero mean and variance σ^2 and coherent demodulation, the recovered signal becomes $y_I(t) = I(t) + n_c$ in the I channel. Because the I and Q channels are symmetrical, we need only consider $y_I(t)$ at $(0, 4T)$ when the sampling time is $2T \pm \Delta t$ where Δt represents the timing jitter, or

$$y_I(2T \pm \Delta t) = I(2T \pm \Delta t) + n_c \quad (19.81)$$

where

$$I(2T \pm \Delta t) = A [a_{-1}g(4T \pm \Delta t) + a_0g(2T \pm \Delta t) + a_1g(\pm \Delta t)] \quad (19.82)$$

Without losing generality, we take the positive sign in (19.81), thus $g(4T + \Delta t) = 0$ and $g(\Delta t) \neq 0$. Therefore (19.81) can be rewritten as

$$y_I(2T + \Delta t) = A[a_0g(2T + \Delta t) + a_1g(\Delta t)] + n_c \quad (19.83)$$

It can be shown that the average power for a signal given by (19.53) is $P_{av} = 2A^2\mu$, where μ is the power of a single pulse $\mu = \frac{1}{4T} \int_0^{4T} g^2(t)dt$. The signal to noise ratio is $r = \frac{2A^2\mu}{\sigma^2}$. After calculating the statistical average over a_0 and a_1 in (19.83), the BER can be shown to be

$$\begin{aligned} P_e &= \frac{1}{4} \operatorname{erfc} \left(\frac{1}{2} \sqrt{\frac{r}{u}} [g(2T + \Delta t) + g(\Delta t)] \right) \\ &+ \frac{1}{4} \operatorname{erfc} \left(\frac{1}{2} \sqrt{\frac{r}{u}} [g(2T + \Delta t) - g(\Delta t)] \right) \end{aligned} \quad (19.84)$$

where $\operatorname{erfc}(x)$ is the complementary error function of x . The BER's for QORC and RCTC modulations with $\Delta t = 0.01T$, $0.25T$ and $0.5T$ are plotted in Figure (19.24). Figure (19.25) shows BER against timing jitter Δt for both modulations with r as a parameter. From Figure (19.24), it is seen that RCTC outperforms QORC in BER for different Δt with the SNR gain ranging from 1.5 dB for $\Delta t = 0.5T$ to 2.2 dB for $t = 0.01T$, for $P_e = 10^{-6}$. Figure (19.25) shows four pairs of curves for $r = 0, 5, 10$ and 15 dB , respectively. When r is large, the performance gap between the two modulations also becomes large. However, if Δt becomes very large, such as $0.5T$, the BER for different signals deteriorates rapidly and the BER gap tends to shrink.

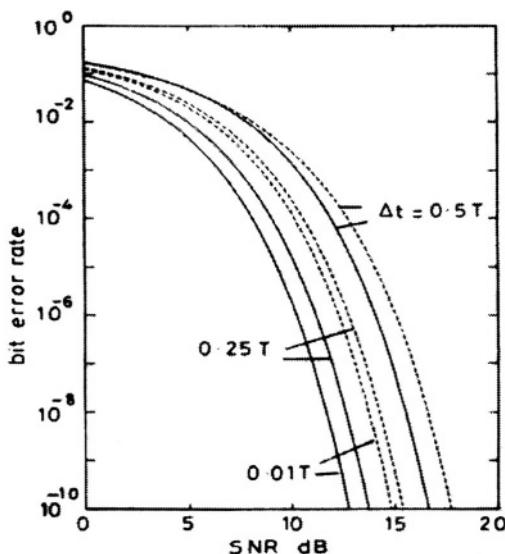


Figure 19.24. Bit error rate for QORC and RCTC modulations under sampling timing jitter Δt , where dotted lines are for QORC and solid lines are for RCTC.

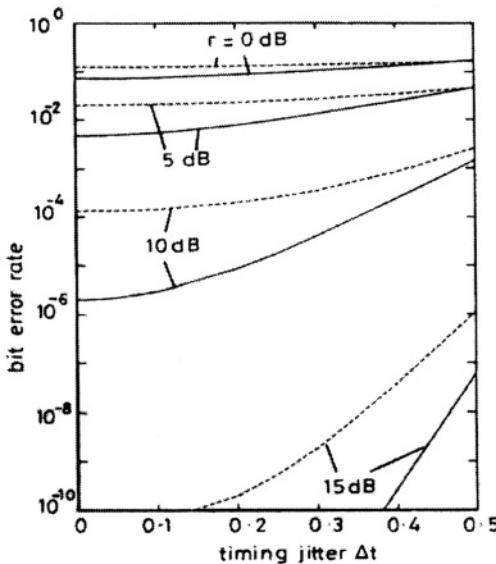


Figure 19.25. Bit error rate for QORC and RCTC modulations with respect to sampling timing jitter Δt , where dotted lines are for QORC and solid lines are for RCTC.

We can summarize from the above analysis about the RCTC modulation and its BER performance under timing jitter as follows. The proposed RCTC modulation performs much better than reported QORC modulation [17] in noisy channels with timing jitter at the receiver. The tolerance of RCTC modulation to timing jitter makes it a suitable modulation scheme for some undesirable working environments, where severe interference may be present to make bit synchronization difficult to track perfectly. It can also be shown that many other pulse-shaped modulations proposed in this chapter have also similar feature, being extremely robust against synchronization timing jitter.

7. Summary

In this chapter we have addressed the issues on pulse-shaping waveform design for bandwidth efficient digital modulations. The chapter starts with a review of the history of symbol waveform design techniques under the context of digital modulations with a high bandwidth efficiency. The importance of the pulse shaping technology and its role in bandwidth efficient digital modems is discussed in section 1. Three major pulse shaping waveform design methodologies have been discussed

in section 2, which include intuitive methods, piece-wise fitting methods and time-domain convolution (TDC) methods. The intuitive methods were widely used in the early research on digital modulations, such as QPSK, OQPSK, MSK, GMSK and QORC, etc. The method is primitive in nature but useful in improving the bandwidth efficiency of a digital modulation scheme. The piece-wise fitting methods employ segmented curves to fit a particular waveform as an effort to enhance the overall spectral characteristics of a digital modem. However, both intuitive methods and piece-wise fitting methods are non-systematic approaches for pulse shaping waveform design, all of which usually involve a very time-consuming try-and-error process to obtain a useful pulse shape for a particular modulation scheme. On the other hand, the TDC method offers several important advantages over the previous two methods in terms of its flexibility and performance. The method itself offers a great degree-of-freedom to control several important spectral parameters of obtained signal, such as the PSD side lobe reduction rate, the width of the PSD main lobe, the time-domain appearance of the pulse-shape, etc. Thus, it is a very promising pulse shaping waveform design method, which can be found very useful in many applications of digital modulation optimization as well as newly emerging UWB technology, where pulse shape design is also of significance in determining the bandwidth efficiency.

In section 3, we have introduced various merit parameters for a pulse-shaping waveform, such as envelope fluctuation ratio ρ , the average power of the PSW μ , the quarter sample point level $g(T)$ and etc., which are closely associated with its bandwidth efficiency in different applications.

Section 4 of the chapter has demonstrated how to analyze bandwidth efficiency of a digital modem under bandpass nonlinear channels, which can always be found in many practical applications, such as satellite systems and mobile cellular transceivers. We have taken the four new QO modulations as examples to show how effectively they could improve their bandwidth efficiency in bandpass nonlinear channels, if compared with conventional modems, such as OQPSK, MSK and QORC, etc. Finally, we have also addressed the issues on robustness against sampling timing jitter for pulse-shaped digital modems in section 5.

Throughout this chapter very much attention has been paid to quadrature-overlapped (QO) modulations due to their great advantages in improving the overall bandwidth-efficiency of a digital modem, which has become an increasingly important issue recently as wireless applications prevail around the world. Several new QO modulation schemes have been proposed in this chapter based on the new pulse-shaping waveform design

method, namely the TDC method, which can be used to replace many traditional digital modulation schemes for their excellent bandwidth efficiency and other attractive characteristics.

Acknowledgments

The author wishes to thank the Editor of this book, Prof. Mohsen Guizani, for his encouragement and support in writing this chapter. He would like also to thank Mr. Yu-Hsin Lin, Mr. Chien-Yao Chao and Mr. Zhe-Min Lin, Institute of Communications Engineering, National Sun Yat-Sen University, Taiwan, for their help in preparing some figures and coding programs for some analysis presented in this chapter. The author would like also to thank National Science Council (NSC), Taiwan, for the following research grants, NSC 89-2213-E-005-032, NSC 89-2213-E-005-029, NSC 89-2213-E-005-032, NSC 90-2213-E-110-054, NSC 91-2213-E-110-032, NSC 92-2213-E-110-015, NSC 92-2213-E-110-047, NSC 92-2213-E-110-048 and NSC 93-2213-E-110-012, which have partially supported the related research work.

References

- [1] Baker, P. A., (1962) Phase Modulation Data Sets for Serial Transmission at 2000 and 2400 bits per second, Part I, *AIEE Trans on Communications Electronics*, July, 1962.
- [2] Bussgang, J. & Leiter, M., (1964) Error rate approximations for differential phase-shift keying, *IEEE Trans on Communications Systems*, vol. CS-12, pp. 18-27, Mar 1964.
- [3] Bussgang, J. & Leiter, (1965) Error performance of a quadrature phase shift keying system, The MITRE Corp., Bedford, Mass., *MITRE Tech. Memo.* 04236, May 14, 1965.
- [4] Bussgang, J. & Leiter, (1966) Phase-shift keying with a transmitted reference, *IEEE Trans. Communication Technology*, vol. COM-14, pp. 14-22, February 1966.
- [5] Bussgang, J. & Leiter, M., (1968) Error Performance of Quadrature Pilot Tone Phase-Shift Keying, *IEEE Transactions on Communications*, pp.526-529, vol: 16, Issue: 4, Aug 1968.
- [6] R. DeBuda, (1972) Coherent demodulation of frequency-shift keying with low deviation ratio, *IEEE Transactions on Communications*. vol. COM-20, pp. 429-435, June 1972.

- [7] S. A. Rhodes, (1972) Effects of hard limiting on bandlimited transmissions with conventional and offset QPSK modulation, *Conf. Rec. of 1972 IEEE National Telecommunications Conf.* 1972, CHO601-5-NTC, pp. 2OF-1, Dec. 1972.
- [8] Mathwich, H., Balcewicz, J. & Hecht, M., (1974) The Effect of Tandem Band and Amplitude Limiting on the Eb/No Performance of Minimum (Frequency) Shift Keying (MSK), *IEEE Transactions on Communications*, pp. 1525-1540, Vol: 22, Issue: 10, Oct 1974.
- [9] Gronemeyer, S. & McBride, A. (1976) MSK and Offset QPSK Modulation, *IEEE Transactions on Communications*, pp.809-820, Vol: 24, Issue: 8, Aug 1976.
- [10] Vasant K. Prabhu, (1977) PSK-type modulation with overlapped baseband pulses, *IEEE Transactions on Communications*, vol. COM-25, No. 9, pp. 980-990, Sept. 1977.
- [11] Oetting, J., (1979) A Comparison of Modulation Techniques for Digital Radio, *IEEE Transactions on Communications*, page(s): 1752-1762, Vol: 27, Issue: 12, Dec 1979.
- [12] Morais, D. & Feher, K., (1979) Bandwidth Efficiency and Probability of Error Performance of MSK and Offset QPSK Systems, *IEEE Transactions on Communications*, pp. 1794-1801, Vol: 27, Issue: 12, Dec 1979.
- [13] F. Amoroso, (1979) The use of quasi-bandlimited pulses in MSK transmission, *IEEE Transactions on Communications*, vol. COM-27, No. 10, pp. 1616-1624, Oct. 1979.
- [14] Korn, I., (1980) Generalized minimum shift keying, *IEEE Transactions on Information Theory*, pp. 234-238, Vol: 26, Issue: 2, Mar 1980.
- [15] D. H. Morais & K. Feher, (1980) The effects of filtering and limiting on the performance of QPSK, offset QPSK, and MSK systems, *IEEE Transactions on Communications*, vol. COM-28, No. 12, pp. 1999-2009, Dec. 1980.
- [16] Murota, K. & Hirade, K., (1981) GMSK Modulation for Digital Mobile Radio Telephony, *IEEE Transactions on Communications*, pp. 1044-1050, Vol: 29, Issue: 7, Jul 1981.
- [17] Mark C. Austin & Ming U. Chang, (1981) Quadrature overlapped raised-cosine modulation, *IEEE Transactions on Communications*, vol. COM-29, No. 3, pp. 237-249, March 1981.

- [18] Milstein, L., Pickholtz, R. & Schilling, D., (1982) Comparison of Performance of Digital Modulation Techniques in the Presence of Adjacent Channel Interference, *IEEE Transactions on Communications*, pp. 1984-1993, Vol: 30, Issue: 8, Aug 1982.
- [19] D. Divsalar & M. K. Simon, (1982) The power spectral density of digital modulations transmitted over nonlinear channels, *IEEE Transactions on Communications*, vol. COM-30, No. 1, pp. 142-151, Jan. 1982.
- [20] A. Yongacoglu & K. Feher, (1985) DCTPSK: An efficient modulation technique for differential detection, *Proceedings of the IEEE ICC*, June, 1985.
- [21] Ali, A.A. & Al-Kadi, I. (1988) A comparison of multipath effect of digital modulations for mobile channels, 1988. *Conference Proceedings on Area Communication*, EUROCON 88., 8th European Conference on Electrotechnics, Stockholm, Sweden, pp. 291-294, 13-17 Jun 1988.
- [22] Kim, K. & Polydoros, A., (1988) Digital modulation classification: the BPSK versus QPSK case, 1988. *MILCOM 88*, Conference record, pp. 431-436, vol.223-26, Oct 1988.
- [23] Cohen, D.J., (1988) Necessary bandwidth of digital modulation, *Symposium Record of IEEE 1988 International Symposium on Electromagnetic Compatibility*, page(s): 247-251, Seattle, USA, 2-4 Aug 1988.
- [24] D. S-Dias & K. Feher, (1991) Baseband pulse shaping techniques for $\pi/4$ -DQPSK in non-linearly amplified land mobile channels, *Proceedings of INFOCOM 1991*, pp. 759-764, 1991.
- [25] S. B. Slimane & T. Le-Ngoc, (1991) Performance of quadrature pulse-overlapping modulated signals in Rayleigh fading channels, *Record of GLOBECOM 1991*, pp. 51.3.1-51.3.5, 1991.
- [26] M. E. Fox & M. W. Marcellin, (1991) Shaped BPSK and the 5 KHz UHF SatCom channels, *Record of MILCOM 1991*, pp. 15.2.1-15.2.7.
- [27] H. H. Chen & J. Oksman, (1992) MSKRC quadrature-overlapping modulation and its performance analysis, *Record of the 3rd international symposium on personal, indoor and mobile radio communications* (PIMRC 1992), pp. 605-609, Boston, USA, October 19-21, 1992.

- [28] S. B. Slimane & T. Le-Ngoc, (1992) Maximum Likelihood sequence estimation of quadrature pulse-overlapping modulated signals for portable/mobile satellite communications, *IEEE Journal on Selected Areas in Communications*, Vol. 10, No. 8, October 1992, pp. 1278-1288.
- [29] H. H. Chen, (1993) Quadrature-overlapped convoluted raised-cosine (QOCRC) modulation and its BER analysis over nonlinear satellite channels, *Record of 1993 international symposium on communications* (ISCOM 1993), National Chiao Tung University, Hsinchu, Taiwan, Vol. 2, pp. 21-28/35, 7-10 December, 1993.
- [30] H. H. Chen, (1993) A novel quadrature-overlapped modulation scheme with immunity to timing jitters at coherent receiver, *Record of 1993 international symposium on personal communications* (ISPC 1993), Nanjing, China, pp. C4. 1 - C4. 6, October 26-28, 1993.
- [31] H. H. Chen & FVC Mendis, (1993) A new bandwidth-efficient digital modulation scheme-MSKRC modulation and its performance analysis in band limited nonlinear channels, *Record of the 5th international symposium on IC technology, system & applications* (ISIC-93), Nanyang Technological University, Singapore, pp. 500-504, 15-17 September 1993.
- [32] H. H. Chen & J. Oksman, (1993) Performance of quadrature-overlapped convoluted triangular-cosine (QOCTC) modulation over nonlinear land mobile satellite channels, *Record of the 4th international symposium on personal, indoor and mobile radio communications* (PIMRC 1993), YoKohama, Japan, pp. 444-447, September 9-11, 1993.
- [33] H. H. Chen, (1993) Convolved raised-cosine and triangular-cosine (RCTC) modulation and its performance against timing jitters over noisy channels, *IEE Electronics Letters*, Vol. 29, No. 14, pp. 1239-1240, 8th July 1993.
- [34] Klovsky, D., Periyalwar, S.S. & Fleisher, S. (1993) Comments on multi-carrier and single-carrier digital modulation in a multipath radio channel, *Canadian Conference on Electrical and Computer Engineering*, Vancouver, Canada, pp. 381-384, vol.1, 14-17 Sep 1993.
- [35] H. H. Chen & J. Oksman, (1994) A robust quadrature-overlapped modulation scheme & its performance over nonlinear band-limited channels, *Record of GLOBECOM 1994*, San Francisco, US, pp. 431-436, Vol. II, November 27-December 1, 1994.

- [36] H. H. Chen & W. H. Ong, (1994) Performance analysis of quadrature overlapping modulations in non-linear fading channels, *Proceedings of ICCS 1994*, pp- 691-695, Vol. II, Singapore, November 14-18, 1994.
- [37] H. H. Chen & J. Oksman, (1996) A Quasi-constant envelope quadrature overlapped modulation & its performance over nonlinear bandlimited satellite channels, *International Journal of Satellite Communications*, John Wiley & Sons, Vol. 14, pp. 351-359, 1996.
- [38] H. H. Chen, (1996) Quadrature overlapping modulation system, *US Patent, No. US5509033*, Patent granted date: April 1996.
- [39] H.H. Chen & S.Y. Wong, Four novel quadrature overlapping modulations and their spectral efficiency analysis over band-limited non-linear channels, *International Journal of Satellite Communications*, John Wiley & Sons, vol. 15, pp. 117-127, 1997.
- [40] H. H. Chen & X. D. Cai, (1997) Optimization of transmitter and receiver filters for the OQAM-OFDM systems by using nonlinear programming algorithms, *IEICE Trans. on Communications*, vol. E80-B, No. 11, November 1997.
- [41] H. H. Chen & S. Y. Wong, (1997) Spectral efficiency analysis of new quadrature overlapped modulations over band-limited non-linear channels, *International Journal of Communication Systems*, John Wiley & Sons, vol. 10, 1-11, pp.1-11, 1997.
- [42] H. H. Chen & X. D. Cai, (1997) Waveform optimization for OQAM-OFDM systems by using nonlinear programming algorithm, *Proceedings of VTC*, 1997, Phoenix, USA, June 1997.
- [43] Parr, B.; ByungLok Cho; Wallace, K.; Zhi Ding; (2003) A novel ultra-wideband pulse design algorithm, *Communications Letters*, IEEE, Volume: 7 Issue: 5, Page(s): 219 -221, May 2003.
- [44] Cheng Yonghong; Chen Xiaolin; Jiang Yan; Lu Liro; Xie Hengkun; (2001) A fingerprint analyzing method of ultra-wideband partial discharge time-domain waveform, *Electrical Insulating Materials*, 2001. (ISEIM 2001). *Proceedings of 2001 International Symposium on*, Page(s): 107-110, 19-22 Nov. 2001.
- [45] Jeongwoo Han; Cam Nguyen; (2002) A new ultra-wideband, ultra-short monocycle pulse generator with reduced ringing, *Microwave and Wireless Components Letters*, IEEE [see also IEEE Microwave and Guided Wave Letters], Volume: 12 Issue: 6, Page(s): 206 -208, June 2002.

- [46] Yue Yansheng; Long Teng; (2000) An ultra-wide band orthogonal digital signal generator, *Signal Processing Proceedings, 2000. WCCC-ICSP 2000. 5th International Conference on*, Volume: 3, Page(s): 2086-2090 vol.3, 21-25 Aug. 2000.
- [47] Xiaojing Huang; Yunxin Li; (2001) Generating near-white ultra-wideband signals with period extended PN sequences, *Vehicular Technology Conference, 2001. VTC 2001 Spring. IEEE VTS 53rd*, Volume: 2, Page(s): 1184 -1188 vol.2, 6-9 May 2001.
- [48] Mesyats, G.A.; Rukin, S.N.; Shpak, V.G.; Yalandin, M.I.; (1998) Generation of high-power subnanosecond pulses, *Ultra-Wideband Short-Pulse Electromagnetics 4*, 1998, Page(s): 1-9, 14-19 June 1998.
- [49] Smithson, G., (1998) Introduction to digital modulation schemes, *IEE Colloquium on The Design of Digital Cellular Handsets*, London , UK, page(s): 2.1-2.9, 4 Mar 1998.
- [50] Hill, F. & Won Lee, (1974) PAM Pulse Generation Using Binary Transversal Filters, *IEEE Transactions on Communications*, page(s): 904-913, Vol: 22, Issue: 7, Jul 1974.
- [51] Civanlar, M.R. & Nobakht, R.A. (1988) Optimal pulse shape design using projections onto convex sets, *1988 International Conference on Acoustics, Speech, and Signal Processing*, page(s): 1874-1877 vol.3, New York, USA, 11-14 Apr 1988.
- [52] Koch, D.B., (1990) Analysis of a generalized coding/decoding method using FIR digital filters for radar waveform design, *Twenty-Second Southeastern Symposium on System Theory*, page(s): 652-656, Cookeville,USA, 11-13 Mar 1990.
- [53] Dzung, D. & Braun, W.R. (1990) Generalized CPM and DPM: digital angle modulation schemes with improved bandwidth efficiency, *IEEE Transactions on Communications*, page(s): 1971-1979, Vol: 38, Issue: 11, Nov 1990.
- [54] Sollenberger, N.R., (1990) Pulse design and efficient generation circuits for linear TDMA modulation, *1990 IEEE 40th Vehicular Technology Conference*, page(s): 616-621, Orlando, USA, 6-9 May 1990.
- [55] Salmasi, A. & Gilhousen, K.S., (1991) On the system design aspects of code division multiple access (CDMA) applied to digital cellular and personal communications networks, *41st IEEE Vehicular Technology Conference, 1991, 'Gateway to the Future Technology in Motion'*, page(s): 57-62, St. Louis, USA, 19-22 May 1991.

- [56] Madhusudhana Rao, G., Reddy, K.V.V.S. & Reddy, E.M.D.E.A. (1993) Waveform design of FH-BFSK communication system useful for mobile telephone subscribers, *Proceedings of the IEEE 1993 National Aerospace and Electronics Conference*, Dayton, USA, page(s): 447-453 vol.1, 24-28 May 1993.
- [57] Stott, G.F., (1994) Digital modulation for radar jamming, *IEE Colloquium on Signal Processing in Electronic Warfare*, London , page(s): 1/1-1/6, UK, 31 Jan 1994.
- [58] Tortoli, P. Guidi, F. & Atzeni, C. (1994) Digital vs. SAW matched filter implementation for radar pulse compression, *Proceedings of 1994 IEEE Ultrasonics Symposium*, Cannes, France, 1-4 Nov 1994.
- [59] Schock, S.G., Leblanc, L.R. & Panda, S. (1994) Spatial and temporal pulse design considerations for a marine sediment classification sonar, *IEEE Journal of Oceanic Engineering*, page(s): 406-415, Vol: 19, Issue: 3, Jul 1994.
- [60] Nikookar, H. & Prasad, R. (1997) Optimal waveform design for multicarrier transmission through a multipath channel, *1997 IEEE 47th Vehicular Technology Conference*, page(s): 1812-1816, vol.3, Phoenix, USA, 4-7 May 1997.
- [61] Chan, Y.T., Plews, J.W. & Ho, K.C. (1997) Symbol rate estimation by the wavelet transform, *Proceedings of 1997 IEEE International Symposium on Circuits and Systems*, Hong Kong, page(s): 177-180 vol.1, 9-12 Jun 1997.
- [62] Lallo, P.R.U., (1999) Signal classification by discrete Fourier transform, *1999 IEEE Military Communications Conference Proceedings*, page(s): 197-201 vol.1, Atlantic City, USA, Oct 31 - Nov 3, 1999.
- [63] Cvetkovic, Z., (1999) Modulating waveforms for OFDM, *Proceedings of 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing*, page(s): 2463-2466 vol.5, Phoenix, USA, Mar 15-19 1999.
- [64] Wern-Ho Sheen, Chun-Chieh Tseng & Chia-Shu Wang, (2000) On the diversity, bandwidth, and performance of digital transmission over frequency-selective slow fading channels, *IEEE Transactions on Vehicular Technology*, page(s): 835-843, Vol: 49, Issue: 3, May 2000.
- [65] Xinjun Zhang, Xin Jiang, Wentao Song & Hanwen Luo (2002) A novel direct waveform synthesis technique with carrier frequency programmable, *2002 IEEE Wireless Communications and Networking Conference*, page(s): 150- 154 vol.1, 17-21 Mar 2002.

Chapter 20

MULTIPLE ANTENNAS

Ezio Biglieri

Politecnico di Torino, Italy

biglieri@polito.it

Giorgio Taricco

Politecnico di Torino, Italy

taricco@polito.it

Abstract In this chapter we investigate transmission systems where more than one antenna is used at both ends of the radio link. The use of multiple transmit and receive antennas allows one to reach capacities that cannot be obtained with any other technique using present-day technology. After computing these capacities, we show how “space–time” codes can be designed, and how suboptimum architectures can be employed to simplify the receiver.

Keywords: Multiple antennas, Fading channels, Space–time codes

1. Preliminaries

The use of multiple antennas for transmission and reception is a major technological advance that is expected to make it possible to increase the data rate in wireless networks by orders of magnitude. While the use of more than one antenna at the receiver side is quite common, only recently was a communication theory of multiple transmit and receive antenna developed. This chapter focuses on the main aspects of this theory.

Here we consider transmission with t antennas and reception with r antennas (Fig. 20.1). Assuming two-dimensional (i.e., complex) elementary constellations throughout, the channel model becomes

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{z} \tag{20.1}$$

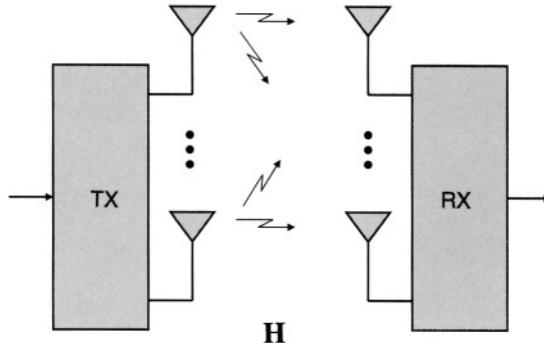


Figure 20.1. Transmission and reception with multiple antennas. The channel gains are described by the $r \times t$ matrix \mathbf{H} .

where $\mathbf{x} \in \mathbb{C}^t$, $\mathbf{y} \in \mathbb{C}^r$, $\mathbf{H} \in \mathbb{C}^{r \times t}$ (i.e., \mathbf{H} is an $r \times t$ complex matrix whose entries h_{ij} describe the gains of each transmission path to a receive from a transmit antenna), and \mathbf{z} is circularly-symmetric complex Gaussian noise. We also assume that

$$\mathbb{E}[\mathbf{z}\mathbf{z}^\dagger] = N_0\mathbf{I}_r \quad (20.2)$$

that is, the noise components affecting the different receivers are independent with variance N_0 , and the signal energy is constrained by $\mathbb{E}[\mathbf{x}^\dagger\mathbf{x}] = tE_s$. Assuming that $\mathbb{E}[|h_{i,j}|^2] = 1$ for all i, j , the average signal-to-noise ratio (SNR) at receiver is

$$\rho \triangleq \frac{tE_s}{N_0}$$

where we assume that the symbol rate equals the signal bandwidth at the Nyquist signalling rate.

The i th component x_i , $i = 1, \dots, t$, of vector \mathbf{x} is the signal transmitted from antenna i ; the j th component y_j , $j = 1, \dots, r$, of vector \mathbf{y} is the signal received by antenna j . Explicitly, we have from (20.1):

$$y_j = \sum_{i=1}^t h_{ij}x_i + z_j, \quad j = 1, \dots, r \quad (20.3)$$

which shows how every component of the received signal includes a linear combination of the signals emitted by each antenna. We say that \mathbf{y} is affected by *spatial interference* from the signals transmitted by the various antennas. This interference has to be removed, or controlled in some way, in order to sort out the transmitted signals. We shall see in the following how this can be done: for the moment we may just observe that the problem of removing interference

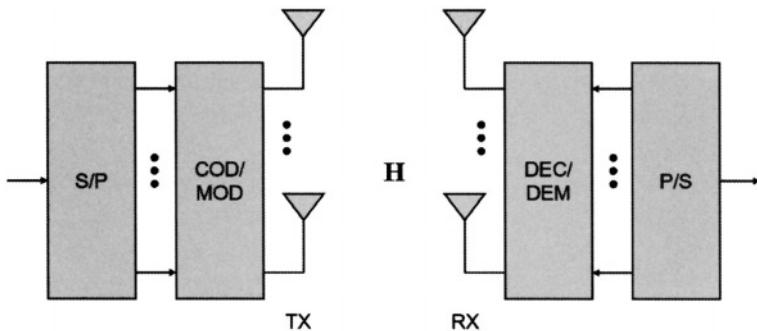


Figure 20.2. Spatial multiplexing through transmission and reception with multiple antennas.

makes the tools for the analysis of multiple-antenna transmission have much in common with those used in the study of digital equalization (which deals with intersymbol interference) and of multiuser detection (which deals with multiple-access interference).

The upsides of using multiple antennas can be summarized by defining two types of gain. As we shall see in the following, in the presence of fading a multiplicity of transmit antennas creates a set of parallel channels, that can be used to potentially increase the data rate up to a factor of $\min\{t, r\}$ (with respect to single-antenna transmission) and hence generate a *rate gain*. This corresponds to the *spatial multiplexing* depicted in Fig. 20.2. Here the serial-to-parallel S/P converter distributes the stream of data across the transmit antennas; after reception, the original stream is reconstituted by the parallel-to-serial converter P/S. The other gain, called *diversity gain*, is due to the combination of received signals that are independently faded replicas of a single transmitted signal, which allows a more reliable reception. We hasten to observe here that these two gains are not independent, but there is a fundamental tradeoff between the two [ZheTse03]. Actually, it can be said that the problem of designing a multiple-antenna systems is based on this tradeoff. As an example, Fig. 20.3 illustrates the diversity–rate tradeoff for a MIMO system with $t = 2$ transmit and $r = 2$ receive antennas. Fig. (a) assumes the channels are orthogonal so that the rate is maximum (twice as much as each channel rate) but there is no diversity gain. Fig. (b) assumes the transmitter replicates the same signal over the two channels so that there is no rate gain but the diversity increases to four.

The problems we address in this chapter are the following:

- What is the limiting performance (channel capacity) of this multiple-antenna system?

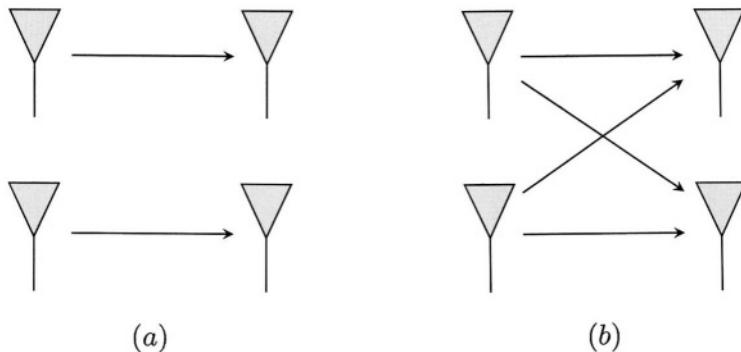


Figure 20.3. Spatial multiplexing through transmission and reception with multiple antennas.

- How can we design codes matched to the channel structure (*space–time codes*)?
- How can we design architectures allowing simple decoding of space–time codes, and what is their performance?

2. Channel capacity

In this section we evaluate the capacity of a multiple-antenna, or MIMO (multiple-input, multiple-output) transmission system as described by (20.1). Unless otherwise specified, capacity will be always measured in bit/dimension pair, corresponding to bit/s/Hz at the Nyquist signalling rate. Several models for the matrix \mathbf{H} can be considered:

- (a) \mathbf{H} is deterministic.
- (b) \mathbf{H} is a random matrix with a given joint probability distribution of its entries, and each channel use (viz., the transmission of one symbol from each of the t transmit antennas) corresponds to an *independent* realization of \mathbf{H} (ergodic channel).
- (c) \mathbf{H} is a random matrix, but once it is chosen it remains fixed for the whole transmission.

When \mathbf{H} is random (cases (b) and (c) above) we assume here that its entries are $\sim \mathcal{N}_c(0, 1)$, i.e., independent, identically distributed (iid) Gaussian with zero-mean, independent real and imaginary parts, each with variance 1/2. Equivalently, each entry of \mathbf{H} has uniform phase and Rayleigh magnitude. This choice models frequency-flat Rayleigh fading with enough separation within

antennas such that the fades for each TX/RX antenna pair are independent. We also assume, for most of our treatment, that the channel state information (CSI, that is, the realization of \mathbf{H}) is known at the receiver, while the distribution of \mathbf{H} is known at the transmitter (the latter assumption is necessary for capacity computations, since the transmitter must choose an optimum code for that specific channel).

Model (c) is applicable to an indoor wireless data network or a personal communication system with mobile terminals moving at walking speed, so that the channel gain, albeit random, varies so slowly with time that it can be assumed as constant along transmission of a long block of data (see [ChuTseKah98,Foschini96,FosGan98,RalCio98,TarSesCal98,Telatar99]). More generally, fading blocks can be thought of as separated in time (e.g., in a time-division system [OzaShaWyn94]), as separated in frequency (e.g., in a multicarrier system), or as separated both in time and in frequency (e.g., with slow time-frequency hopping [CaiKnoHum98, Knopp97, KnoHum00]). This model, the block-fading (BF) channel [BigProSha98], allows the incorporation of delay constraints, which imply that, even though very long code words are transmitted, perfect (i.e., infinite-depth) interleaving cannot be achieved. With this fading model, the channel may or may not be ergodic (ergodicity is lost when there is a delay constraint). Without ergodicity, the average mutual information over the ensemble of channel realizations cannot characterize the achievable transmission rates. Since the instantaneous mutual information of the channel turns out to be a random variable, the quantity that characterizes the quality of a channel when coding is used is not channel capacity, but rather the *information outage probability*, i.e., the probability that this mutual information is lower than the rate of the code used for transmission [OzaShaWyn94]. This outage probability is closely related to the code word error probability, as averaged over the random coding ensemble and over all channel realizations; hence, it provides useful insight on the performance of a delay-limited coded system [CaiKnoHum98, Knopp97, KnoHum00, MalLei97, MalLei99]. An additional important definition related to outage probability is that of *zero-outage capacity*, sometimes also referred to as *delay-limited capacity*. This is the maximum rate for which the minimum outage probability is zero under a given power constraint [BigCaiTar01,CaiTarBig98,CaiTarBig99,TseHanly98].

2.1 Deterministic channel

Capacity is the maximum of the average mutual information $I(\mathbf{x}; \mathbf{y})$ between input and output of the channel over the choice of the distribution of \mathbf{x} .

Singular-value decomposition (SVD) of matrix \mathbf{H} yields

$$\mathbf{H} = \mathbf{U}\mathbf{D}\mathbf{V}^\dagger \quad (20.4)$$

where $\mathbf{U} \in \mathbb{C}^{r \times r}$ and $\mathbf{V} \in \mathbb{C}^{t \times t}$ are unitary, and $\mathbf{D} \in \mathbb{R}^{r \times t}$ is diagonal. We can write

$$\mathbf{y} = \mathbf{UDV}^\dagger \mathbf{x} + \mathbf{z} \quad (20.5)$$

Premultiplication of (20.5) by \mathbf{U}^\dagger shows that the original channel is equivalent to the channel described by the input-output relationship

$$\tilde{\mathbf{y}} = \mathbf{D}\tilde{\mathbf{x}} + \tilde{\mathbf{z}} \quad (20.6)$$

where $\tilde{\mathbf{y}} \triangleq \mathbf{U}^\dagger \mathbf{y}$, $\tilde{\mathbf{x}} \triangleq \mathbf{V}^\dagger \mathbf{x}$ (so that $\mathbb{E}[\tilde{\mathbf{x}}^\dagger \tilde{\mathbf{x}}] = \mathbb{E}[\mathbf{x}^\dagger \mathbf{x}]$), and $\tilde{\mathbf{z}} \triangleq \mathbf{U}^\dagger \mathbf{z} \sim \mathcal{N}_c(0, N_0 \mathbf{I}_r)$. Now, the rank of \mathbf{H} is at most $m \triangleq \min\{t, r\}$, and hence at most m of its singular values are nonzero. Denote these by $\sqrt{\lambda_i}$, $i = 1, \dots, m$, and rewrite (20.6) componentwise in the form

$$\tilde{y}_i = \sqrt{\lambda_i} \tilde{x}_i + \tilde{z}_i, \quad i = 1, \dots, m \quad (20.7)$$

which corresponds to a set of m independent parallel additive Gaussian noise channels. The remaining components of $\tilde{\mathbf{y}}$ (if any) are equal to the corresponding components of the noise vector $\tilde{\mathbf{z}}$: we see that, for $i > m$, \tilde{y}_i is *independent of the transmitted signal*, and \tilde{x}_i does not play any role.

Defining ρ_i the SNR constraint for the i th channel, the overall channel capacity is

$$C = \sum_{i=1}^m \log(1 + \lambda_i \rho_i) \quad (20.8)$$

Using the constraint $\sum_i \rho_i \leq \rho$ we get the “water-filling” solution [CovTho91, Gal68]

$$\rho_i(\mu) = (\mu - \lambda_i^{-1})_+ \quad (20.9)$$

where μ is obtained by solving the equation

$$\rho = \sum_{i=1}^m (\mu - \lambda_i^{-1})_+ \quad (20.10)$$

The capacity can then be written as

$$C = \sum_{i=1}^m (\log(\mu \lambda_i))_+ \quad (20.11)$$

and is attained by choosing $\tilde{x}_i \sim \mathcal{N}_c(0, N_0(\mu - \lambda_i^{-1})_+)$ for $i = 1, \dots, m$, i.e., iid circularly-symmetric zero-mean complex Gaussian random variables with variances $N_0(\mu - \lambda_i^{-1})_+$.

Example 1. Consider as \mathbf{H} the all-1 matrix. The SVD of \mathbf{H} yields

$$\mathbf{H} = \mathbf{UDV}^\dagger = \begin{bmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \dots & 1 \end{bmatrix} = \frac{1}{\sqrt{r}} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} [\sqrt{rt}] \frac{1}{\sqrt{t}} [1, \dots, 1]$$

Since there is only one singular value we have the trivial solution $\rho_1 = \rho$ yielding the channel capacity

$$C = \log(1 + rt\rho) \quad (20.12)$$

The signals achieving this capacity can be described as follows. Since $\tilde{\mathbf{x}}_1 \sim \mathcal{N}_c(0, \rho N_0)$, the vector $\mathbf{x} = \mathbf{V}\tilde{\mathbf{x}}_1$ has all equal components $\sim \mathcal{N}_c(0, \rho N_0/t)$. Thus, the TX antennas all send the same signal with power $\rho N_0/t$. Because of the structure of \mathbf{H} the signals *add coherently* at the receiver, so that at each receiver we have the voltage $t\sqrt{\rho N_0/t}$, and hence the power $t^2 \times \rho N_0/t = \rho N_0 t$. Since each receiver sees the same signal, and the noises are uncorrelated, the overall SNR is $rt\rho$, as shown by the capacity formula (20.12). \square

Example 2. Take $t = r = m$, and $\mathbf{H} = \mathbf{I}_m$. Due to the structure of \mathbf{H} , there is no spatial interference here, and transmission occurs over m parallel AWGN channels, each having SNR ρ/m and hence capacity $\log(1 + \rho/m)$. Thus,

$$C = m \log(1 + \rho/m) \quad (20.13)$$

Notice how the capacity here increases faster with the number of antennas than in Example 1: this is due to the absence of spatial interference. Notice also that as m grows to infinity the capacity converges to the finite limit $\rho \log e$. \square

2.2 Ergodic Rayleigh fading channel

We assume here that \mathbf{H} is independent of both \mathbf{x} and \mathbf{z} , with entries $\sim \mathcal{N}_c(0, 1)$. We also assume that for each channel use an *independent* realization of \mathbf{H} is drawn, so that the channel is memoryless. If the receiver has perfect knowledge of the realization of \mathbf{H} (the “channel state information,” or CSI), the channel capacity is achieved by a transmitted signal $\mathbf{x} \sim \mathcal{N}_c(0, (\rho/t)\mathbf{I}_t)$, and takes value [Telatar99]

$$C = \mathbb{E} \left[\log \det \left(\mathbf{I}_r + \frac{\rho}{t} \mathbf{H} \mathbf{H}^\dagger \right) \right] \quad (20.14)$$

Note that for fixed r and as $t \rightarrow \infty$, the strong law of large numbers yields

$$\frac{1}{t} \mathbf{H} \mathbf{H}^\dagger \rightarrow \mathbf{I}_r \quad \text{a.s.} \quad (20.15)$$

Thus, as $t \rightarrow \infty$ the capacity tends almost surely to

$$\log \det (\mathbf{I}_r + \rho \mathbf{I}_r) = r \log(1 + \rho) \quad (20.16)$$

so that the capacity increases *linearly* with r . This result should be compared with (20.12), where C increases only *logarithmically* with r .

The expectation in (20.14) can be computed explicitly, yielding [Telatar99]

$$C = \int_0^\infty \log \left(1 + \frac{\rho}{t} \lambda \right) \sum_{k=0}^{m-1} \frac{k!}{(k-n+m)!} [L_k^{n-m}(\lambda)]^2 \lambda^{n-m} e^{-\lambda} d\lambda \quad (20.17)$$

where $m \triangleq \min\{t, r\}$, $n \triangleq \max\{t, r\}$, and $L_k^\ell(\lambda)$ is the associated Laguerre polynomial defined as [Szegő39]:

$$L_k^\ell(\lambda) \triangleq \frac{1}{k!} e^x x^{-\ell} \frac{d^k}{dx^k} [e^{-x} x^{k+\ell}] \quad (20.18)$$

Capacity values for $\rho = 20$ dB are plotted in Fig. 20.4 and 20.5. Some special cases as well as asymptotic approximations to the values of C are examined in the examples that follow.

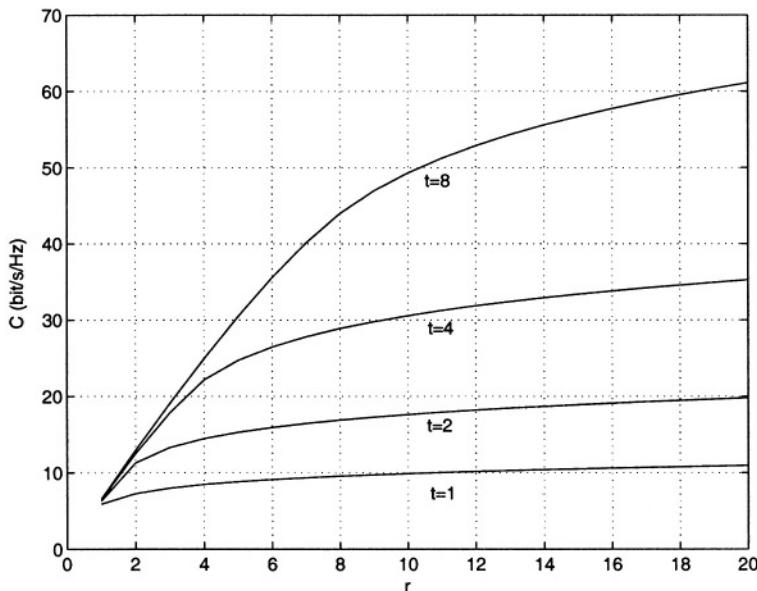


Figure 20.4. Capacity of the ergodic Rayleigh MIMO channel with $\rho = 20$ dB.

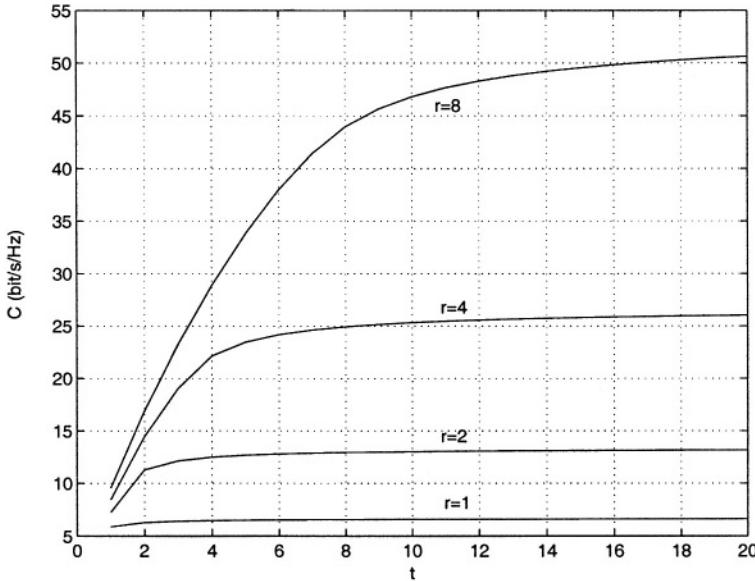


Figure 20.5. Capacity of the ergodic Rayleigh MIMO channel with $\rho = 20$ dB.

Example 3 ($t = 1$, $r \rightarrow \infty$). Consider first the case $t = 1$, so that $m = 1$ and $n = r$. Since by definition $L_0^{n-m}(\cdot) = 1$, application of (20.17) yields

$$C = \frac{1}{(r-1)!} \int_0^\infty \log(1 + \rho\lambda) \lambda^{r-1} e^{-\lambda} d\lambda$$

This is plotted in Fig. 20.6. An asymptotic expression of C as $r \rightarrow \infty$ is $C \sim \log(1 + \rho r)$. This approximation to the capacity is also plotted in Fig. 20.6. We see here that if $t = 1$ the capacity increases only logarithmically as the number of receive antennas is increased—a quite inefficient way of boosting capacity. \square

Example 4 ($r = 1$, $t \rightarrow \infty$). Consider first $r = 1$, so that $m = 1$ and $n = t$. Application of (20.17) yields

$$C = \frac{1}{(t-1)!} \int_0^\infty \log\left(1 + \frac{\rho}{t}\lambda\right) \lambda^{t-1} e^{-\lambda} d\lambda$$

This is plotted in Fig. 20.7. An asymptotic expression of C as $t \rightarrow \infty$ is $C \sim \log(1 + \rho)$. This approximation to the capacity is also plotted in Fig. 20.7. \square

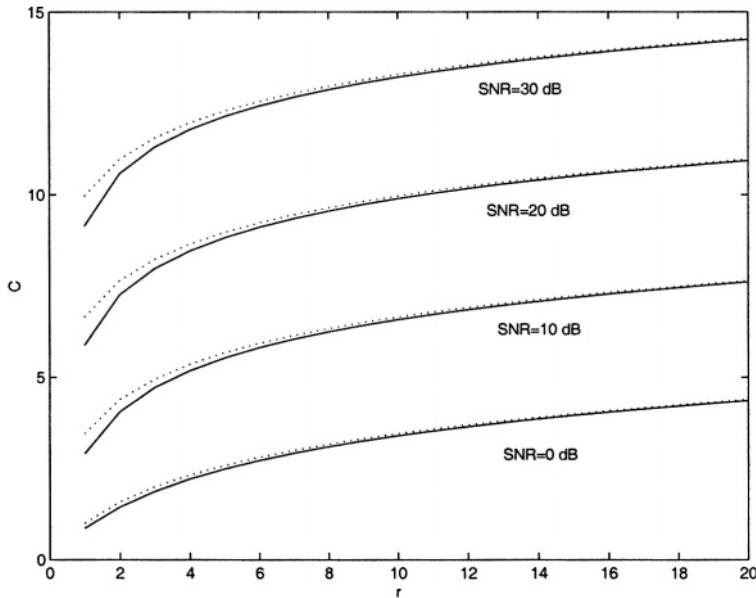


Figure 20.6. Capacity of the ergodic Rayleigh MIMO channel with $t = 1$ (continuous line). The asymptotic approximation $C \sim \log(1 + \rho r)$ is also shown (dotted line).

Example 5 ($r = t$). With $r = t$ we have $m = n = r$, so that application of (20.17) yields

$$C = \int_0^\infty \log\left(1 + \frac{\rho}{r}\lambda\right) \sum_{k=0}^{r-1} [L_k(\lambda)]^2 e^{-\lambda} d\lambda$$

where $L_k \triangleq L_k^0$ is a Laguerre polynomial of order k (see eq. (20.18)). The capacity is plotted in Fig. 20.8. \square

A general asymptotic approximation to (20.14) can be derived by using results from random-matrix theory [Sil95]. We have

$$\frac{C}{m} \rightarrow \left(\log(w_+ + \rho) + (1 - \alpha) \log(1 - w_-) - \frac{w_- - \alpha}{\ln 2} \right) \max(1, 1/\alpha) \quad (20.19)$$

where

$$w_\pm \triangleq (w \pm \sqrt{w^2 - 4/\alpha})/2 \quad (20.20)$$

and

$$w \triangleq 1 + \frac{1}{\alpha} + \frac{1}{\rho} \quad (20.21)$$

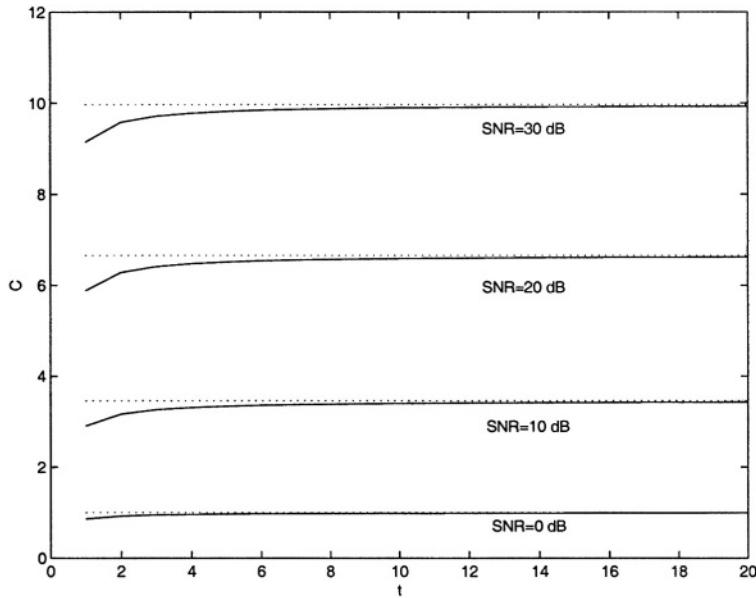


Figure 20.7. Capacity of the ergodic Rayleigh MIMO channel with $r = 1$ (continuous line). The asymptotic approximation $C \sim \log(1 + \rho)$ is also shown (dotted line).

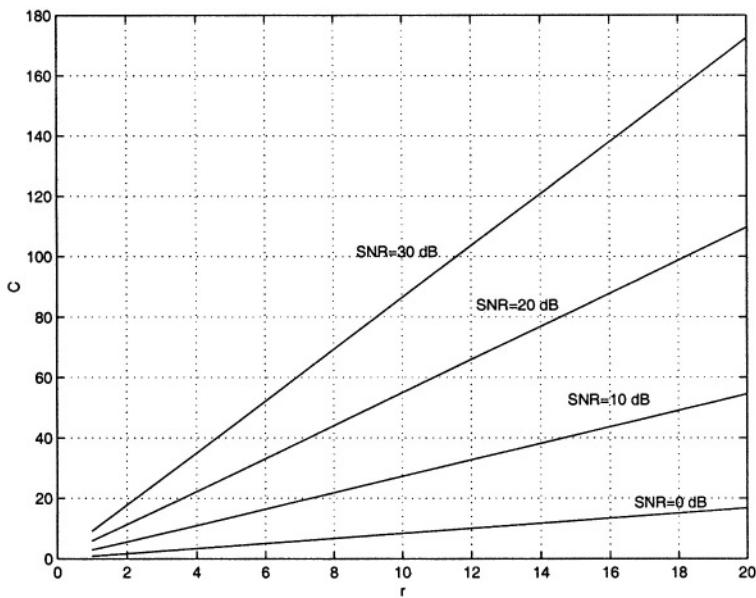


Figure 20.8. Capacity with independent Rayleigh fading and $r = t$ antennas.

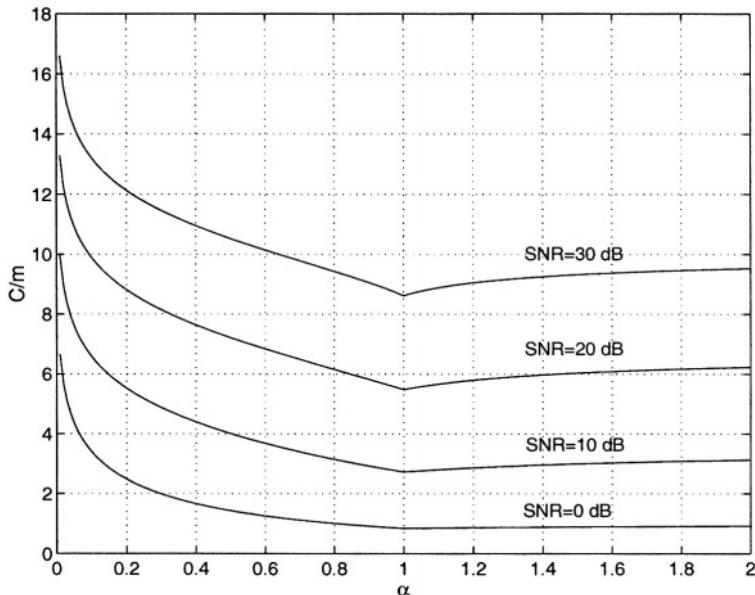


Figure 20.9. Capacity per antenna as $t, r \rightarrow \infty$ with independent Rayleigh fading and $t/r \rightarrow \alpha$.

This asymptotic result (plotted in Fig. 20.9) can be used to approximate the value of C for finite r, t .

The Appendix shows a simple way to derive the above result. By setting $\alpha = t/r$, (20.19) yields C as a function of t and r , and provides values very close to the true capacity even for small r, t . Tables 20.1–20.3 show some examples of exact and asymptotic values of capacity for different values of α and of the SNR ρ .

Table 20.1. Exact values of C/m for $t/r = 1/3$ and different values of SNR and of t .

| ρ (dB) | $t = 1$ | $t = 2$ | $t = 4$ | $t = 8$ | $t \rightarrow \infty$ |
|-------------|---------|---------|---------|---------|------------------------|
| 0 dB | 1.87 | 1.87 | 1.86 | 1.86 | 1.86 |
| 10 dB | 4.72 | 4.71 | 4.70 | 4.70 | 4.70 |
| 20 dB | 7.98 | 7.97 | 7.96 | 7.96 | 7.96 |
| 30 dB | 11.30 | 11.28 | 11.28 | 11.28 | 11.28 |

Table 20.2. Exact values of C/m for $t/r = 1/2$ and different values of SNR and of t .

| ρ (dB) | $t = 1$ | $t = 2$ | $t = 4$ | $t = 8$ | $t \rightarrow \infty$ |
|-------------|---------|---------|---------|---------|------------------------|
| 0 dB | 1.44 | 1.43 | 1.43 | 1.43 | 1.43 |
| 10 dB | 4.06 | 4.02 | 4.01 | 4.01 | 4.01 |
| 20 dB | 7.27 | 7.23 | 7.22 | 7.22 | 7.22 |
| 30 dB | 10.58 | 10.54 | 10.53 | 10.53 | 10.53 |

Table 20.3. Exact values of C/m for $t/r = 2$ and different values of SNR and of t .

| ρ (dB) | $t = 2$ | $t = 4$ | $t = 8$ | $t = 16$ | $t \rightarrow \infty$ |
|-------------|---------|---------|---------|----------|------------------------|
| 0 dB | 0.92 | 0.92 | 0.91 | 0.91 | 0.91 |
| 10 dB | 3.17 | 3.14 | 3.13 | 3.13 | 3.13 |
| 20 dB | 6.28 | 6.24 | 6.23 | 6.23 | 6.23 |
| 30 dB | 9.58 | 9.54 | 9.53 | 9.53 | 9.53 |

Observation 1. For the validity of (20.19), it is not necessary to assume that the entries of \mathbf{H} are Gaussian: a sufficient condition is that \mathbf{H} have iid entries with unit variance.

Observation 2. If $C(r, t, \rho)$ denotes the capacity of a channel with r receive antennas, t transmit antennas, and total SNR ρ , we have

$$C(a, b, \rho b) = C(b, a, \rho a) \quad (20.22)$$

Thus, for example, $C(r, 1, \rho) = C(1, r, r\rho)$, which shows that with TX rather than RX diversity we need r times as much transmit power to achieve the same capacity.

Observation 3. Choose $t = r = 1$ as the baseline; this yields *one* more bit/s/Hz for every 3 dB of SNR increase. In fact, for large ρ ,

$$C = \log(1 + \rho) \sim \log \rho \quad (20.23)$$

For multiple antennas with $t = r$, for every 3 dB of SNR increase we have t more bit/s/Hz.

A critique. The previous results derived under the assumption $r \rightarrow \infty$ should be accepted *cum grano salis*. The assumption that the entries of the channel-gain matrix \mathbf{H} are independent random variables becomes increasingly questionable as r increases. In fact, for this assumption to be justified

the antennas should be separated by some multiple of the wavelength, which cannot be obtained when a large number of antennas is packed in a finite volume. Furthermore, if the variance of the elements of \mathbf{H} does not depend on r , increasing r leads to an increased total received power, which is physically unacceptable. These observations lead to the consideration of a multiple-antenna system where not only the total transmit power remains constant as t increases, but also the average received power remains constant when r increases. This is obtained by rescaling \mathbf{H} by a factor $r^{-1/2}$, so that the capacity (20.14) becomes

$$C = \mathbb{E} \left[\log \det \left(\mathbf{I}_r + \frac{\rho}{rt} \mathbf{H} \mathbf{H}^\dagger \right) \right] \quad (20.24)$$

As t and r grows to infinity, we obtain [ChiRimTel01]

$$C \rightarrow \rho \log e \quad (20.25)$$

a conclusion in contrast with our previous result that capacity increases linearly with the number of antennas.

2.3 Nonergodic Rayleigh fading channel

When \mathbf{H} is chosen randomly at the beginning of the transmission, and held fixed for all channel uses, average capacity has no meaning, because the channel is nonergodic. In this case the quantity to be evaluated is, rather than capacity, outage probability. Under these conditions, the mutual information of the channel is the random variable

$$C(\mathbf{H}) = \log \det \left(\mathbf{I}_r + \frac{\rho}{t} \mathbf{H} \mathbf{H}^\dagger \right) \quad (20.26)$$

and the outage probability is given by

$$P_{\text{out}}(R) = \mathbb{P}(C(\mathbf{H}) < R) \quad (20.27)$$

The evaluation of (20.27) can be done by Monte Carlo simulation. However, one can profitably use an asymptotic result from matrix theory which states that, as t and r grow to infinity, the capacity $C(\mathbf{H})$ tends to a Gaussian random variable. This result has been recently obtained independently by a variety of authors under slightly different assumptions [HocMarTar02, SenMit02, SmiSha02].

The expected value of $C(\mathbf{H})$ is the ergodic capacity discussed in previous section. Its variance was evaluated independently in the form of an integral for finite r and t in [SmiSha02] and [WanGia02] (the latter reference derives the moment generating function of $C(\mathbf{H})$, and hence all of its moments). Moreover, it was observed in [SmiSha02] that $C(\mathbf{H})$ follows very closely a normal distribution even for small values of r and t . Thus, if μ_C and σ_C^2 denote the

asymptotic mean and variance of $C(\mathbf{H})$, respectively, then a close approximation of the outage probability can be obtained in the form

$$P_{\text{out}}(R) \approx Q\left(\frac{\mu_C - R}{\sigma_C}\right) \quad (20.28)$$

where $Q(\cdot)$ is the Gaussian tail function. Refs. [SenMit02, MouSimSen03] derive results that can be specialized to our problem to yield

$$\sigma_C^2 = -\log e \log\left(1 - \frac{\gamma^2 \delta^2}{\beta}\right) \quad (20.29)$$

with $\beta \triangleq \alpha^{-1}$, and γ, δ are solutions of the equation pair

$$\gamma = \frac{\beta}{w + \delta} \quad \delta = \frac{1}{w + \gamma} \quad (20.30)$$

with $w \triangleq \sqrt{1/\rho}$, which in our case yields

$$\gamma \triangleq \frac{\beta - 1 - w^2 + \sqrt{(\beta - 1 - w^2)^2 + 4\beta w^2}}{2w} \quad (20.31)$$

$$\delta \triangleq \frac{1 - \beta - w^2 + \sqrt{(1 - \beta - w^2)^2 + 4w^2}}{2w} \quad (20.32)$$

An alternative expression in the form of an integral for the asymptotic capacity variance was obtained in [BaiSil02]. This integral can be evaluated numerically, and results are shown in Fig. 20.10; the limiting value of the variance as $\rho \rightarrow \infty$ is

$$\lim_{\rho \rightarrow \infty} \sigma_C^2 = \log e \cdot \log(1 - \min(\alpha, \alpha^{-1})) \quad (20.33)$$

Fig. 20.11, which plots P_{out} versus R for $r = t = 4$ and two values of SNR, shows the quality of the Gaussian approximation for a Rayleigh channel.

Based on these results, we can evaluate the outage probabilities as in Figs. 20.12 and 20.13. These figures show the rate R that can be supported by the channel for a given SNR and a given outage probability, that is, from (20.28):

$$R = \mu_C - \sigma_C Q^{-1}(P_{\text{out}})$$

Notice how as r, t increase the outage probabilities curves come closer to each other: in fact, as r and t grow to infinity the channel tends to an ergodic channel. Finally, Fig. 20.14 shows the outage capacity (at $P_{\text{out}} = 0.01$) of an independent Rayleigh fading MIMO channel.

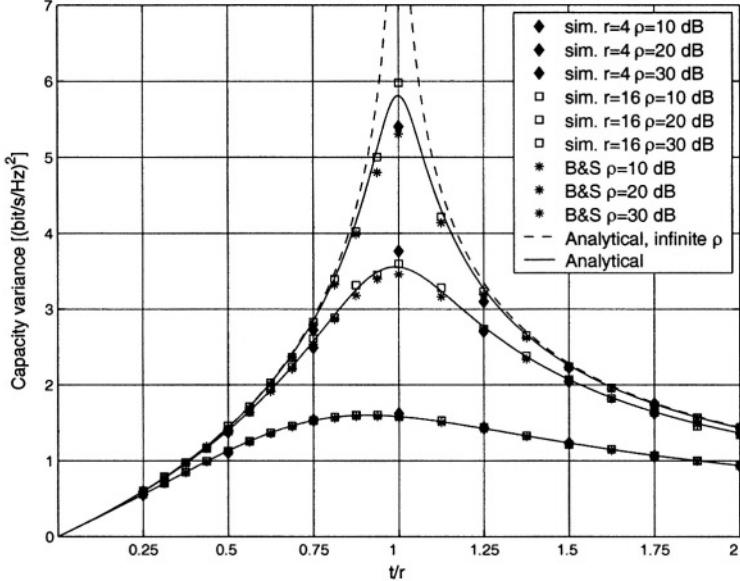


Figure 20.10. Variance of the nonergodic capacity. Continuous line: Asymptotic variance from [SenMit02, MouSimSen03]. (*): Numerical integration of Bai-Silverstein expression for the variance [BaiSil02]. (♦) and (□): Monte Carlo simulation.

Block-fading channel. We now choose a block-fading channel model as shown in Fig. 20.15. The channel is described by the F matrices \mathbf{H}_k , $k = 1, \dots, F$, each describing the fading gains in a single block. These gains are independent of each other. The channel input-output equation is

$$\mathbf{y}_k[\ell] = \mathbf{H}_k \mathbf{x}_k[\ell] + \mathbf{z}_k[\ell] \quad (20.34)$$

for $k = 1, \dots, F$ (block index) and $\ell = 1, \dots, N$ (symbol index along a block), $\mathbf{y}_k, \mathbf{z}_k \in \mathbb{C}^r$, and $\mathbf{x}_k \in \mathbb{C}^t$.

It is convenient to use the singular-value decomposition

$$\mathbf{H}_k = \mathbf{U}_k \mathbf{D}_k \mathbf{V}_k^\dagger \quad (20.35)$$

where \mathbf{D}_k is a $r \times t$ matrix whose main-diagonal elements are the ordered singular values $\sqrt{\lambda_{k,1}} \geq \dots \geq \sqrt{\lambda_{k,m}}$, with $\lambda_{k,i}$ the i th largest eigenvalue of the Hermitian matrix $\mathbf{H}_k \mathbf{H}_k^\dagger$, and $m \triangleq \min(r, t)$. Since \mathbf{U}_k and \mathbf{V}_k are unitary, by premultiplying $\mathbf{y}_k[\ell]$ by \mathbf{U}_k^\dagger the input-output relation (20.34) can be rewritten in the form

$$\tilde{\mathbf{y}}_k[\ell] = \mathbf{D}_k \tilde{\mathbf{x}}_k[\ell] + \tilde{\mathbf{z}}_k[\ell] \quad (20.36)$$

where $\tilde{\mathbf{y}}_k[\ell] \triangleq \mathbf{U}_k^\dagger \mathbf{y}_k[\ell]$, $\tilde{\mathbf{x}}_k[\ell] \triangleq \mathbf{V}_k^\dagger \mathbf{x}_k[\ell]$, $\tilde{\mathbf{z}}_k[\ell] \triangleq \mathbf{U}_k^\dagger \mathbf{z}_k[\ell]$, and $\tilde{\mathbf{z}}_k[\ell] \sim \mathbf{z}_k[\ell] \sim \mathcal{N}_c(0, N_0)$.

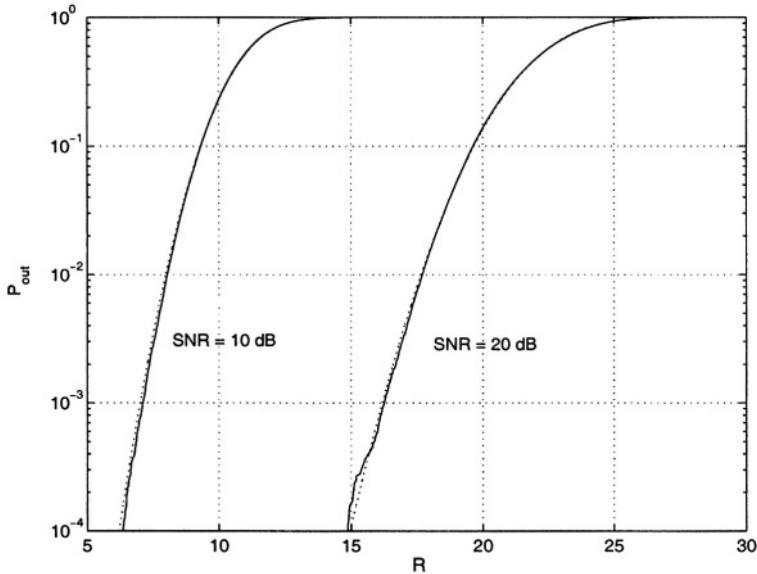


Figure 20.11. Outage probability for $r = t = 4$ and a nonergodic Rayleigh channel vs. R , the transmission rate in bit/s/Hz. The continuous line shows the results obtained by Monte Carlo simulation, while the dashed line shows the normal approximation.

No delay constraints. Since the random matrix process $\{\mathbf{H}_k\}_{k=1}^F$ is iid, as $F \rightarrow \infty$ the channel is ergodic and the average capacity is the relevant quantity. When perfect CSI is available to the receiver only, this is given by

$$C = \sum_{i=1}^m \mathbb{E} \left[\log \left(1 + \frac{\rho}{t} \lambda_i \right) \right] \quad (20.37)$$

If perfect CSI is available to transmitter and receiver, we proceed as in Section 2.1 and obtain

$$C = \mathbb{E} \left[\sum_{i=1}^m (\log(\mu \lambda_i))_+ \right] \quad (20.38)$$

where μ is the solution of the “water-filling” equation

$$\sum_{i=1}^m (\mu - 1/\lambda_i)_+ = \rho \quad (20.39)$$

For all block lengths $N = 1, 2, \dots$, the capacities (20.37) and (20.38) are achieved by code sequences with length $F N t$ with $F \rightarrow \infty$. Capacity (20.37)

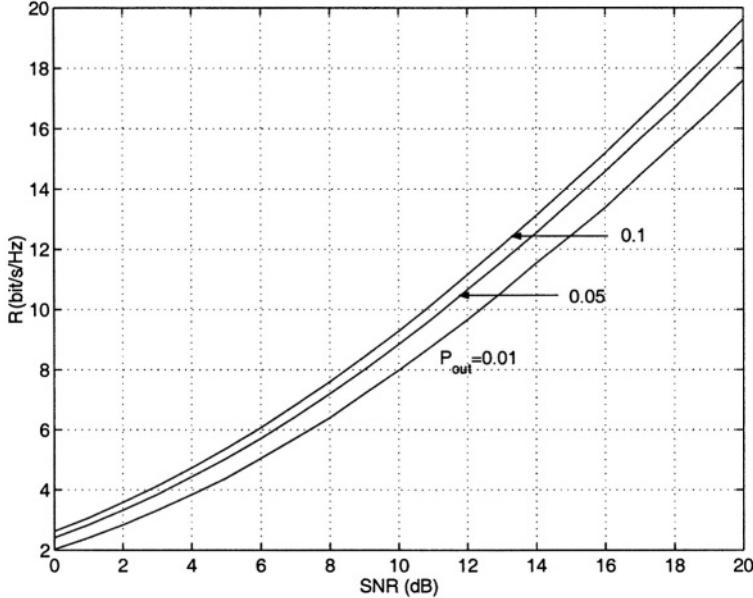


Figure 20.12. Transmission rate that can be supported with $r = t = 4$ and a give outage probability by a nonergodic Rayleigh channel. The results are based on the Gaussian approximation.

is achieved by random codes whose symbols are iid complex $\sim \mathcal{N}_c(0, \rho/t)$. Thus, all antennas transmit the same average energy per symbol. capacity (20.38) can be achieved by generating a random code with iid components $\sim \mathcal{N}_c(0, 1)$ and having each code word split into F blocks of N vectors $\tilde{\mathbf{x}}_k[\ell]$ with t components each. For block k , the optimal linear transformation

$$\mathbf{W}_k = \mathbf{V}_k \operatorname{diag}(\sqrt{\rho_{k,1}}, \dots, \sqrt{\rho_{k,m}}, \underbrace{0, \dots, 0}_{t-m}) \quad (20.40)$$

is computed, where $\rho_{i,k} \triangleq (\xi - 1/\lambda_{k,i})_+$. The vectors $\mathbf{x}_k[\ell] = \mathbf{W}_k \tilde{\mathbf{x}}_k[\ell]$ are transmitted from the t antennas. This optimal scheme can be viewed as the concatenation of an optimal encoder for the unfaded AWGN channel, followed by a beamformer with weighting matrix \mathbf{W}_k varying from block to block [Big-CaiTar01].

Delay constraints. Consider now a delay constraint that forces F to take on a finite value. The random capacity is given by

$$C(\boldsymbol{\Lambda}, \boldsymbol{\Gamma}) = \frac{1}{F} \sum_{k=1}^F \sum_{i=1}^m \log(1 + \lambda_{k,i} \rho_{k,i}) \quad (20.41)$$

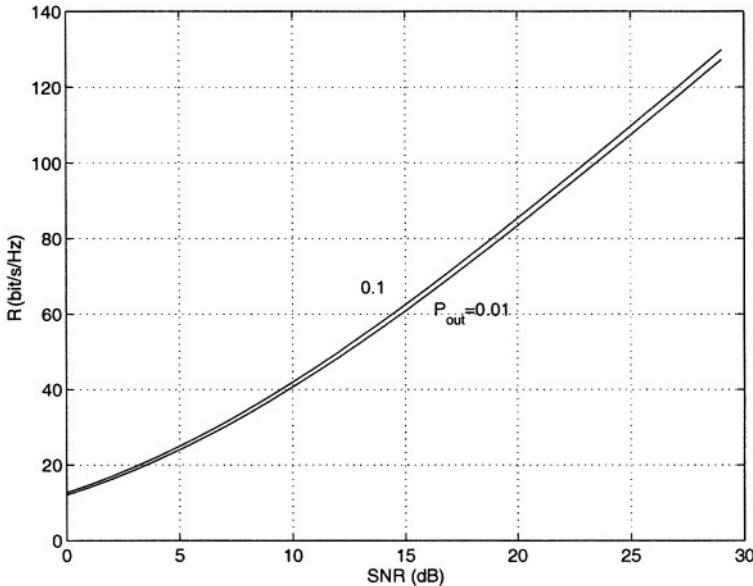


Figure 20.13. Transmission rate that can be supported with $r = t = 16$ and a give outage probability by a nonergodic Rayleigh channel. The results are based on the Gaussian approximation.

which corresponds to the outage probability

$$P_{\text{out}}(R) = \mathbb{P}(C(\mathbf{\Lambda}, \mathbf{\Gamma}) < R) \quad (20.42)$$

The outage probability is minimized under a long-term average power constraint by setting a threshold SNR* such that if the SNR per block necessary to avoid an outage exceeds SNR* then transmission is turned off and an outage is declared, while if it is below SNR* transmission is turned on, and the power is allocated to the blocks according to a rule depending on the fading statistics only through the threshold value SNR* (see [BigCaiTar01]).

2.4 Influence of channel-state information

A crucial factor in determining the performance of a multi-antenna system is the availability of the channel-state information (CSI), that is, the knowledge of the values of the fading gains in each one of the transmission paths. As we have seen, in a system with t transmit and r receive antennas and an ergodic Rayleigh fading channel modeled by an $t \times r$ matrix with random i.i.d. complex Gaussian entries the average channel capacity with perfect CSI at the receiver is about $m \triangleq \min(t, r)$ times larger than that of a single-antenna system for the same transmitted power and bandwidth. The capacity increases by about

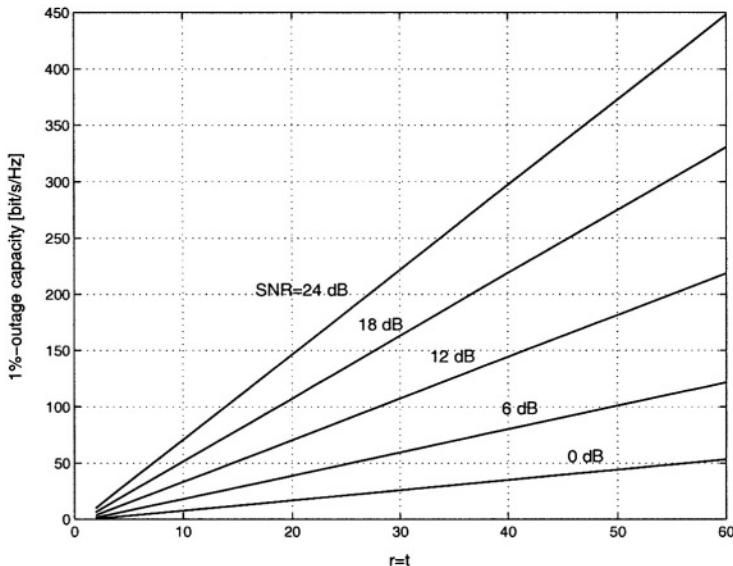


Figure 20.14. Outage capacity (at $P_{\text{out}} = 0.01$) with independent Rayleigh fading and $r = t$ antennas.

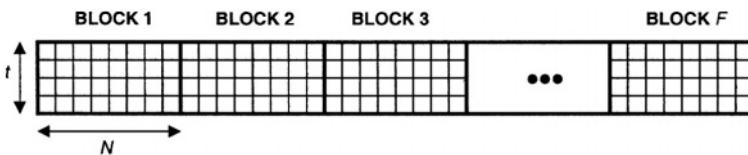


Figure 20.15. One code word in an F -block fading channel.

m bit/s/Hz for every 3-dB increase in SNR. Due to the assumption of perfect CSI available at the receiver, this result can be viewed as a fundamental limit for coherent multiple-antenna systems [ZheTse02].

In a fixed wireless environment, the fading gains can be expected to vary slowly, so their estimate can be obtained by the receiver with a reasonable accuracy, even in a system with a large number of antennas. One way of obtaining this estimate is by periodically sending pilot signals on the same channel used for data signals: if the channel is assumed to remain constant for I symbol periods, then we may write $I = I_r + I_d$, where I_r is the number of pilot symbols, while data transmission occupies I_d symbols. Since the transmission of pilot symbols lowers the information rate, there is a tradeoff between system performance and transmission rate.

Perfect CSI at the receiver. The most commonly studied situation is that of perfect CSI available at the receiver, which is the assumption under which we described multiple-antenna systems above.

No CSI. Fundamental limits of non-coherent communication, i.e., one taking place in an environment where estimates of the fading coefficients are not available, were derived in [MarHoc99, HocMar98]. In this channel model the fading gains are iid Rayleigh and remain constant for I symbol periods before changing to new independent realizations. Under these assumptions, further increasing the number of transmit antennas beyond I cannot increase capacity. Zheng and Tse [ZheTse02] derive an explicit formula for the high-SNR average channel capacity for $t = r$ and $I \geq 2r$, and characterize the rate at which capacity increases with SNR for $I < 2r$. By defining $K \triangleq \min(r, \lfloor I/2 \rfloor)$, [ZheTse02] shows that the capacity gains $K(1 - K/I)$ bits per second per Hz for every 3-dB SNR increase.

Imperfect CSI at the receiver. In real world the receiver has an imperfect knowledge of the CSI. In [TarBig03] some answers are provided to two questions: 1) How long should the training interval I_r be for satisfactory operation? 2) What is the performance degradation caused by imperfect estimation of CSI?

It can be shown that we must have $I_r \geq t$, i.e., the duration of the training interval must be at least as great as the number of transmit antennas. Moreover, the optimum training signals are orthogonal with respect to time among the transmit antennas, and each transmit antenna is fed equal energy.

CSI at the transmitter and at the receiver. It is also possible to envisage a situation in which channel state information is known to the receiver and to the transmitter: the latter can take the appropriate measures to counteract the effect of channel attenuations by suitably modulating its power. To assure causality, the assumption of CSI available at the transmitter is valid if it is applied to a multicarrier transmission scheme in which the available frequency band (over which the fading is selective) is split into a number of subbands, as with OFDM. The subbands are so narrow that fading is frequency-flat in each of them, and they are transmitted simultaneously, via orthogonal subcarriers. From a practical point of view, the transmitter can obtain the CSI either from a dedicated feedback channel (some existing systems already implement a fast power-control feedback channel) or by time-division duplex, where the uplink and the downlink time-share the same subchannels and the fading gains can be estimated from the incoming signal.

Ref. [BigCaiTar01] derives the performance limits of a channel with additive white Gaussian noise, delay and transmit-power constraints, and perfect

channel-state information available at both transmitter and receiver. Because of a delay constraint, the transmission of a code word is assumed to span a finite (and typically small) number of independent channel realizations; therefore, the channel is nonergodic, and the relevant performance limits are the information outage probability and the delay-limited capacity. The coding scheme that minimizes the information outage probability was also derived. This scheme can be interpreted as the concatenation of an optimal code for the AWGN channel without fading to an optimal linear beamformer, whose coefficients change whenever the fading changes. For this optimal scheme we evaluated minimum-outage probability and delay-limited capacity. Among other results, [BigCaiTar01] proves that, for the fairly general class of *regular* fading channels, the asymptotic delay-limited capacity slope, expressed in bit/s/Hz per dB of transmit SNR, is proportional to $m \triangleq \min(t, r)$ and independent of the number of fading blocks F . Since F is a measure of the time diversity (induced by interleaving) or of the frequency diversity of the system, this result shows that, if channel-state information is available also to the transmitter, very high rates with asymptotically small error probabilities are achievable without need of deep interleaving or high frequency diversity. Moreover, for a large number of antennas the delay-limited capacity approaches the ergodic capacity. Finally, the availability of CSI at the transmitter makes transmit-antenna diversity equivalent, in terms of capacity improvement, to receive-antenna diversity, in the sense that reciprocity holds.

3. Coding for multiple-antenna systems

Given that considerable gains are achievable by a multi-antenna system, the challenge is to design coding schemes that perform close to capacity: space-time trellis codes, space-time block codes, and layered space-time codes have been advocated (see, e.g., [GueFitBelKuo99, ShiKah99, ShiKah99a, SteDum99, TarSesCal98, TarJafCal99]).

A space-time code with block length N is described by the $t \times N$ matrix $\mathbf{X} \triangleq (\mathbf{x}[1], \dots, \mathbf{x}[N])$. The code, which we denote \mathfrak{X} , has $|\mathfrak{X}|$ words. The row index of \mathbf{X} indicates space, while the column index indicates time: to wit, the i th component of the t -vector $\mathbf{x}[\ell]$, denoted $x_i[\ell]$, is a complex number representing the two-dimensional signal transmitted by the i th antenna at discrete time ℓ , $\ell = 1, \dots, N$, $i = 1, \dots, t$. The received signal is the $r \times N$ matrix

$$\mathbf{Y} = \mathbf{H}\mathbf{X} + \mathbf{Z} \quad (20.43)$$

where \mathbf{Z} is matrix of zero-mean complex Gaussian random variables (RV) with zero mean and independent real and imaginary parts with the same variance $N_0/2$ (i.e., circularly-distributed). Thus, the noise affecting the received signal is spatially and temporally independent, with $\mathbb{E}[\mathbf{Z}\mathbf{Z}^\dagger] = NN_0\mathbf{I}_r$, where \mathbf{I}_r denotes the $r \times r$ identity matrix and $(\cdot)^\dagger$ denotes Hermitian transposition.

The channel is described by the $r \times t$ matrix \mathbf{H} . Here we assume that \mathbf{H} is independent of both \mathbf{X} and \mathbf{Z} , it remains constant during the transmission of an entire code word, and its realization (the CSI) is known at the receiver.

3.1 Maximum likelihood detection

Under the assumption of CSI perfectly known by the receiver, and of additive white Gaussian noise, maximum-likelihood decoding corresponds to choosing the code word \mathbf{X} which minimizes the squared Frobenius norm $\|\mathbf{Y} - \mathbf{HX}\|^2$. Explicitly, ML detection and decoding corresponds to the minimization of the quantity

$$\|\mathbf{Y} - \mathbf{HX}\|^2 = \sum_{i=1}^r \sum_{n=1}^N \left| y_{in} - \sum_{j=1}^t h_{ij} x_{jn} \right|^2 \quad (20.44)$$

Pairwise error probability. For computations, we may resort to the union bound to error probability

$$P(e) \leq \frac{1}{|\mathcal{X}|} \sum_{\mathbf{X} \in \mathcal{X}} \sum_{\hat{\mathbf{X}} \neq \mathbf{X}} P(\mathbf{X} \rightarrow \hat{\mathbf{X}}) \quad (20.45)$$

where the pairwise error probability (PEP) $P(\mathbf{X} \rightarrow \hat{\mathbf{X}})$ is given by [BenBig99]

$$P(\mathbf{X} \rightarrow \hat{\mathbf{X}}) = \mathbb{E} \left[Q \left(\frac{\|\mathbf{H}\Delta\|}{\sqrt{2N_0}} \right) \right] \quad (20.46)$$

where $\Delta \triangleq \mathbf{X} - \hat{\mathbf{X}}$. By writing

$$\|\mathbf{H}\Delta\|^2 = \text{Tr} \left(\mathbf{H}^\dagger \mathbf{H} \Delta \Delta^\dagger \right) \quad (20.47)$$

we see that the exact pairwise error probability, and hence the union bound to $P(e)$, depends on the $t \times r$ matrix $\mathbf{H}^\dagger \mathbf{H}$. This matrix can be interpreted as representing the effect of spatial interference on error probability: in particular, if $\mathbf{H}^\dagger \mathbf{H} = \mathbf{I}_t$ then (20.46) becomes

$$P(\mathbf{X} \rightarrow \hat{\mathbf{X}}) = Q \left(\frac{\|\Delta\|}{\sqrt{2N_0}} \right) \quad (20.48)$$

which is the PEP we would obtain on a set of t parallel independent AWGN channels, each transmitting a code formed by a row of \mathbf{X} .

A simple and useful upper bound to the pairwise error probability (20.46) can be computed by bounding above the Q function:

$$Q \left(\frac{\|\mathbf{H}\Delta\|}{\sqrt{2N_0}} \right) \leq \exp(-\|\mathbf{H}\Delta\|^2/4N_0) \quad (20.49)$$

Under the assumption of independent Rayleigh fading, that is, when the entries h_{ij} of \mathbf{H} are independent complex Gaussian random variables, circularly distributed with variance of their real and imaginary parts equal to $1/2$, the exact expectation of the RHS of the above can be computed:

$$P(\mathbf{X} \rightarrow \widehat{\mathbf{X}}) \leq \det \left[\mathbf{I}_t + \boldsymbol{\Delta} \boldsymbol{\Delta}^\dagger / 4N_0 \right]^{-r} \quad (20.50)$$

The rank-and-determinant criterion. Since the determinant of a matrix is equal to the product of its eigenvalues, (20.50) yields

$$P(\mathbf{X} \rightarrow \widehat{\mathbf{X}}) = \prod_{j=1}^t (1 + \lambda_j / 4N_0)^{-r} \quad (20.51)$$

where λ_j denotes the j th eigenvalue of $\boldsymbol{\Delta} \boldsymbol{\Delta}^\dagger$. We can also write

$$P(\mathbf{X} \rightarrow \widehat{\mathbf{X}}) \leq \prod_{j \in \mathcal{J}} (\lambda_j / 4N_0)^{-r} \quad (20.52)$$

where \mathcal{J} is the index set of the nonzero eigenvalues of $\boldsymbol{\Delta} \boldsymbol{\Delta}^\dagger$. Denoting by ν the number of elements in \mathcal{J} , and rearranging the indexes so that $\lambda_1, \dots, \lambda_R$ are the nonzero eigenvalues, we have

$$P(\mathbf{X} \rightarrow \widehat{\mathbf{X}}) \leq \left[\prod_{j=1}^{\nu} \lambda_j \right]^{-r} \gamma^{-r\nu} \quad (20.53)$$

where $\gamma \triangleq 1/4N_0$. From this expression we see that the total diversity order (the “diversity gain”) of the coded system is $r\nu_{\min}$, where ν_{\min} is the minimum rank of $\boldsymbol{\Delta} \boldsymbol{\Delta}^\dagger$ across all possible pairs $\mathbf{X}, \widehat{\mathbf{X}}$. In addition, the pairwise error probability depends on the r th power of the product of eigenvalues of $\boldsymbol{\Delta} \boldsymbol{\Delta}^\dagger$. This does not depend on the SNR (which is proportional to γ), and displaces the error probability curve instead of changing its slope. This is called the “coding gain.” Thus, for high enough SNR we can choose, as a criterion for designing a space–time code, the maximization of the coding gain as well as of the diversity gain.

Notice that if $\nu_{\min} = t$, i.e., $\boldsymbol{\Delta} \boldsymbol{\Delta}^\dagger$ is full-rank for all the code word pairs, we have

$$\prod_{j=1}^t \lambda_j = \det(\boldsymbol{\Delta} \boldsymbol{\Delta}^\dagger) \quad (20.54)$$

A necessary condition for $\boldsymbol{\Delta} \boldsymbol{\Delta}^\dagger$ to be full-rank is that $N \geq t$, i.e., the code block length must be at least equal to the number of transmit antennas.

The Euclidean-distance criterion. Observe now that the term in the upper bound (20.50) can be written as

$$\det(\mathbf{I}_t + \gamma \Delta \Delta^\dagger) = 1 + \gamma \text{Tr}(\Delta \Delta^\dagger) + \dots + \gamma^t \det(\Delta \Delta^\dagger) \quad (20.55)$$

where $\Delta \triangleq \mathbf{X} - \widehat{\mathbf{X}}$. We see that if $\gamma \ll 1$ then the LHS of (20.55), and hence the PEP, depends essentially on $\text{Tr}(\Delta \Delta^\dagger)$, which is the Euclidean distance between \mathbf{X} and $\widehat{\mathbf{X}}$, while if $\gamma \gg 1$ it depends essentially on $\det(\Delta \Delta^\dagger)$, that is, on the product of the eigenvalues of $\Delta \Delta^\dagger$.

A different perspective can be obtained by allowing the number r of receive antennas to grow to infinity. To do this, we first renormalize the entries of \mathbf{H} so that their variance is now $1/r$ rather than 1: this prevents the total receive power to diverge as $r \rightarrow \infty$. We obtain the following new form of (20.50):

$$P(\mathbf{X} \rightarrow \widehat{\mathbf{X}}) \leq \det \left[\mathbf{I}_t + \Delta \Delta^\dagger / 4rN_0 \right]^{-r} \quad (20.56)$$

which yields, in lieu of (20.55):

$$\det[\mathbf{I}_t + (\gamma/r) \Delta \Delta^\dagger] = 1 + (\gamma/r) \text{Tr}(\Delta \Delta^\dagger) + \dots + (\gamma/r)^t \det(\Delta \Delta^\dagger) \quad (20.57)$$

This shows that as $r \rightarrow \infty$ the rank-and-determinant criterion is appropriate for a SNR increasing as fast as r , while the Euclidean-distance criterion is appropriate for relatively small SNRs. This situation is illustrated in the example of Fig. 20.16, which shows the union upper bound (obtained numerically) on the word-error probability of the space-time code obtained by splitting evenly the code words of the (24, 8, 12) extended Golay code between two transmit antennas. This space-time code has the minimum rank of Δ equal to 1, and hence a diversity gain equal to r . Now, it is seen from Fig. 20.16 how the slope predicted by (20.50), and exhibited by a linear behavior in the $P(e)$ -vs.- E_b/N_0 chart, may be reached only for very small values of error probability (how small, that depends on the weight structure of the code under scrutiny). To justify this behavior, observe from Fig. 20.16 that for a given value of r the error-probability curve changes its behavior from a “waterfall” shape (for small to intermediate SNR) to a linear shape (high SNR). As the number of receive antennas grows, this change of slope occurs for values of $P(e)$ that are smaller and smaller as r increases. Thus, to study the word-error-probability curve in its waterfall region it makes sense to examine its asymptotic behavior as $r \rightarrow \infty$. The case $r \rightarrow \infty$, $t < \infty$ can be easily dealt with by using the law of large numbers: this yields $\mathbf{H}^\dagger \mathbf{H} \rightarrow \mathbf{I}_t$, \mathbf{I}_t the $t \times t$ identity matrix. We can see that, as $r \rightarrow \infty$,

$$\|\mathbf{H}\Delta\|^2 \rightarrow \|\Delta\|^2 \quad (20.58)$$

and hence

$$P(\mathbf{X} \rightarrow \widehat{\mathbf{X}}) = Q \left(\frac{\|\Delta\|}{\sqrt{2N_0}} \right) \quad (20.59)$$

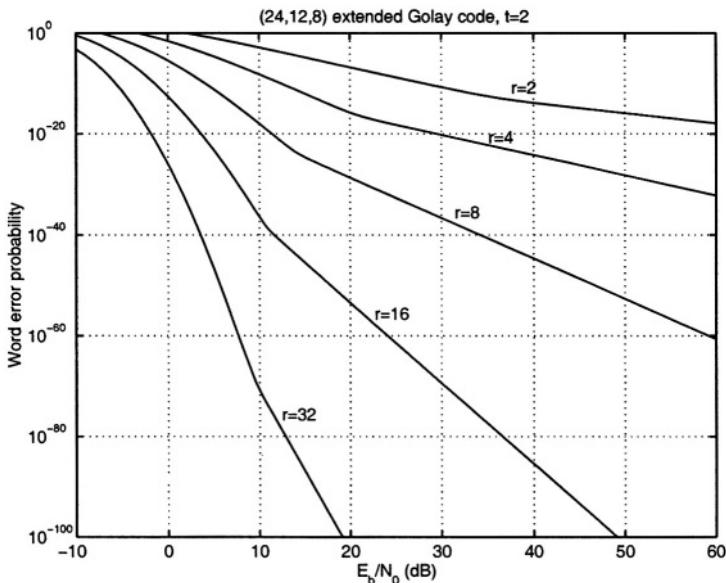


Figure 20.16. Performance of the binary $(24, 8, 12)$ extended Golay code with binary PSK over a channel with $t = 2$ transmit antennas and r receive antennas with ML interface.

This result shows that as the number of receiving antennas grows large the union bound on the error probability of the space-time code depends only on the Euclidean distances between pairs of code words. This is the result one would get with a transmission occurring over a non-fading additive white Gaussian noise (AWGN) channel whose transfer matrix \mathbf{H} has orthogonal columns, i.e., is such that $\mathbf{H}^\dagger \mathbf{H}$ is a scalar matrix. In this situation the smallest error probability, at the expense of a larger complexity, can be achieved by using a single code, optimized for the AWGN channel, whose words of length tN are equally split among the transmit antennas. Within this framework, the number of transmit antennas does not affect the PEP, but only the transmission rate which, expressed in bits per channel use, increases linearly with t .

For another example, observe Fig. 20.17. This shows how for intermediate SNRs the Euclidean criterion may yield codes better than the rank-and-determinant criterion. It compares the simulated performances, in terms of frame-error rate, of the 4-state, rate-1/2 space-time code of [TarLo98] and a comparable space-time code obtained by choosing a good binary, 4-state, rate-2/4 convolutional code [ChaHwaLin97] and mapping its symbols onto QPSK. The frame length N is 130 symbols for both codes, including 1 symbol for trellis termination. It is seen that in the error-probability interval of these two figures the “standard” convolutional code generally outperforms the space-time code of [TarLo98] even for small values of r .

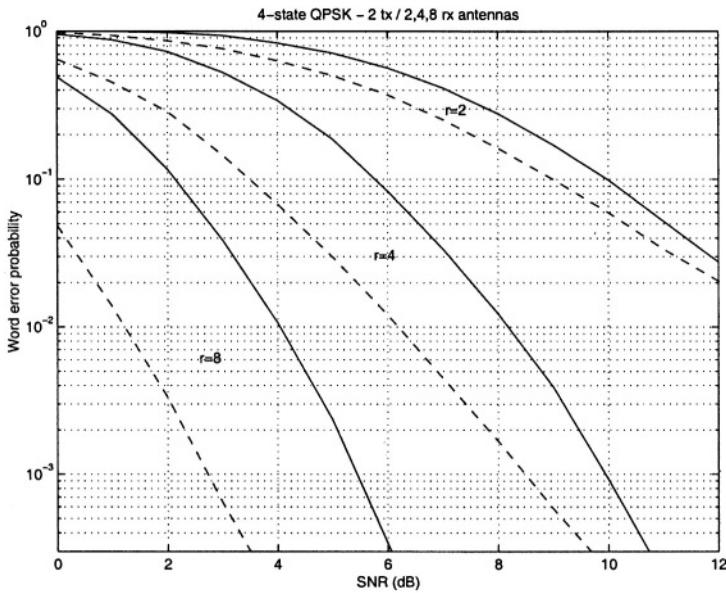


Figure 20.17. Frame-error rates of two space–time codes with 4 states, rate 1/2, and QPSK. Number of transmit antennas: $t = 2$; number of receive antennas: $r = 2, 4, 8$. Continuous line: Code from [TarLo98]. Dashed line: Code obtained from a binary convolutional code good for the AWGN channel [ChaHwaLin97].

Decoding space–time codes. Decoding of space–time codes is performed by using Viterbi algorithm whose branch metrics at time n are, from (20.44):

$$\sum_{i=1}^r \left| y_{in} - \sum_{j=1}^t h_{ij} x_{jn} \right|^2 \quad (20.60)$$

Computation of (20.60) requires knowledge of the fading gains h_{kj} , i.e., CSI at the receiver.

4. Space–time coding schemes

4.1 Alamouti scheme

This allows transmission of one signal per channel use, and removes the spatial interference by orthogonalization. We describe it by first considering the simple case $t = 2, r = 1$, which yields the scheme illustrated in Fig. 20.18.

The code matrix \mathbf{X} has the form

$$\mathbf{X} = \begin{bmatrix} x_1 & -x_2^* \\ x_2 & x_1^* \end{bmatrix} \quad (20.61)$$

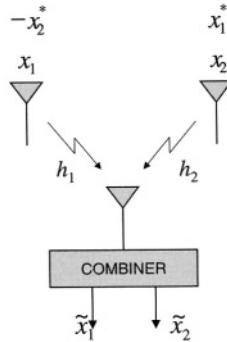


Figure 20.18. Alamouti TX-diversity scheme with $t = 2$ and $r = 1$.

This means that during the first symbol interval, the signal x_1 is transmitted from antenna 1, while signal x_2 is transmitted from antenna 2. During the next symbol period, antenna 1 transmits signal $-x_2^*$, and antenna 2 transmits signal x_1^* . Thus, the signals received in two adjacent time slots are

$$y_1 = h_1 x_1 + h_2 x_2 + z_1$$

and

$$y_2 = -h_1 x_2^* + h_2 x_1^* + z_2$$

where h_1 , h_2 denote the path gains from the two TX antennas to the RX antenna. The combiner of Fig. 20.18, which has perfect CSI and hence knows the values of the path gains, generates the signals

$$\tilde{x}_1 = h_1^* y_1 + h_2^* y_2^*$$

and

$$\tilde{x}_2 = h_2^* y_1 - h_1^* y_2^*$$

so that

$$\begin{aligned} \tilde{x}_1 &= h_1^*(h_1 x_1 + h_2 x_2 + z_1) + h_2(-h_1^* x_2 + h_2^* x_1 + z_2^*) \\ &= (|h_1|^2 + |h_2|^2)x_1 + (h_1^* z_1 + h_2 z_2^*) \end{aligned} \quad (20.62)$$

and similarly

$$\tilde{x}_2 = (|h_1|^2 + |h_2|^2)x_2 + (h_2^* z_1 - h_1 z_2^*) \quad (20.63)$$

Thus, we have separated x_1 from x_2 . This scheme has the same performance as one with $t = 1$, $r = 2$, and maximal-ratio combining (provided that each TX antenna transmits the same power as the single antenna for $t = 1$). To prove the last statement, observe that if the signal x_1 is transmitted, the two

receive antennas observe $h_1x_1 + n_1$ and $h_2x_1 + n_2$, respectively, and after maximal-ratio combining the decision variable is

$$h_1^*(h_1x_1 + z_1) + h_2^*(h_2x_1 + z_2) = (|h_1|^2 + |h_2|^2)x_1 + (h_1^*z_1 + h_2^*z_2^*) = \tilde{x}_1$$

This scheme can be generalized to other values of r . For example, with $t = r = 2$ and the same transmission scheme as before (see Fig. 20.19), we

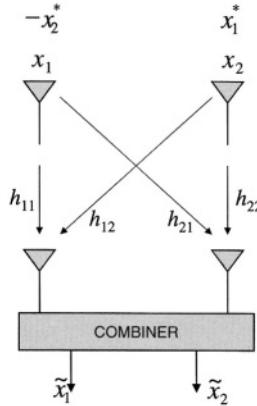


Figure 20.19. Alamouti TX-diversity scheme with $t = 2$ and $r = 2$.

have, if $y_{11}, y_{12}, y_{21}, y_{22}$ denote the signals received by antenna 1 at time 1, by antenna 1 at time 2, by antenna 2 at time 1, and by antenna 2 at time 2, respectively,

$$\begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix} \begin{bmatrix} x_1 & -x_2^* \\ x_2 & x_1^* \end{bmatrix} + \begin{bmatrix} z_{11} & z_{12} \\ z_{21} & z_{22} \end{bmatrix}$$

$$= \begin{bmatrix} h_{11}x_1 + h_{12}x_2 + z_{11} & -h_{11}x_2^* + h_{12}x_1^* + z_{12} \\ h_{21}x_1 + h_{22}x_2 + z_{21} & -h_{21}x_2^* + h_{22}x_1^* + z_{22} \end{bmatrix}$$

The combiner generates

$$\tilde{x}_1 = h_{11}^*y_1 + h_{12}^*y_2 + h_{21}^*y_3 + h_{22}^*y_4$$

and

$$\tilde{x}_2 = h_{12}^*y_1 - h_{11}^*y_2 + h_{22}^*y_3 - h_{21}^*y_4$$

which yields

$$\tilde{x}_1 = (|h_{11}|^2 + |h_{12}|^2 + |h_{21}|^2 + |h_{22}|^2)x_1 + \text{noise}$$

and

$$\tilde{x}_2 = (|h_{11}|^2 + |h_{12}|^2 + |h_{21}|^2 + |h_{22}|^2)x_2 + \text{noise}$$

As above, it can be easily shown that the performance of this $t = 2, r = 2$ scheme is equivalent to that of a $t = 1, r = 4$ scheme with maximal-ratio combining (again, provided that each TX antenna transmits the same power as with $t = 1$).

A general scheme with $t = 2$ and a general value of r can be exhibited: it has the same performance of a single-TX-antenna scheme with $2r$ RX antennas and maximal-ratio combining.

4.2 Alamouti scheme revisited: Orthogonal designs

We can rewrite the transmitted signal in Alamouti scheme with $t = 2$ and $r = 1$ in the following equivalent form:

$$\begin{bmatrix} y_1 \\ y_2^* \end{bmatrix} = \begin{bmatrix} h_1 & h_2 \\ h_2^* & -h_1^* \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \quad (20.64)$$

Now, if we define

$$\check{\mathbf{H}} \triangleq \begin{bmatrix} h_1 & h_2 \\ h_2^* & -h_1^* \end{bmatrix}$$

we see that

$$\check{\mathbf{H}}^\dagger \check{\mathbf{H}} = (|h_1|^2 + |h_2|^2) \mathbf{I}_2 \quad (20.65)$$

This, in conjunction with (20.46) and (20.47), shows that the error probability for this Alamouti scheme is the same as without spatial interference, and with a received power increased by a factor $(|h_1|^2 + |h_2|^2)$. For this reason Alamouti scheme is called an *orthogonal design*. There are also orthogonal designs with $t > 2$. For example, with $t = 3, r = 1$, and $N = 4$ we have

$$\mathbf{X} = \begin{bmatrix} x_1 & -x_2^* & -x_3^* & 0 \\ x_2 & x_1^* & 0 & -x_3^* \\ x_3 & 0 & x_1^* & x_2^* \end{bmatrix}$$

so that the equation $\mathbf{Y} = \mathbf{H}\mathbf{X} + \mathbf{Z}$ can be rewritten in the equivalent form

$$\begin{bmatrix} y_1 \\ y_2^* \\ y_3^* \\ y_4^* \end{bmatrix} = \check{\mathbf{H}} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \quad (20.66)$$

where

$$\check{\mathbf{H}} \triangleq \begin{bmatrix} h_1 & h_2 & h_3 \\ h_2^* & -h_1^* & 0 \\ h_3^* & 0 & h_1^* \\ 0 & h_3^* & -h_2^* \end{bmatrix} \quad (20.67)$$

In this case we can verify that

$$\check{\mathbf{H}}^\dagger \check{\mathbf{H}} = (|h_1|^2 + |h_2|^2 + |h_3|^2) \mathbf{I}_3$$

Notice that with this code we transmit three signals in four time intervals, so that its rate is $3/4$ signals per channel use, less than the original Alamouti schemes, which transmit 1 signal per channel use. In fact, orthogonal designs with $t > 2$ have rates that cannot exceed $3/4$ [WanGia02]. To avoid the rate loss of orthogonal designs, algebraic codes can be designed that, for any number of transmit and receive antennas, achieve maximum diversity, as Alamouti codes, while the rate is t symbols per channel use (see [MaGian02] and references therein).

4.3 Trellis space–time codes

Trellis space–time codes are TCM schemes in which every transition among states, as described by a trellis, is labeled by t signals, each being associated with a transmit antenna [TarSesCal98].

Example 7. An example of a space–time code is shown in Fig. 20.20 through its trellis. This has $t = 2$, four states, and transmits 2 bit/channel use by using 4PSK, whose signals are denoted 0, 1, 2, 3. With $r = 1$ its diversity is 2. Label xy means that signal x is transmitted by antenna 1, while signal y is simultaneously transmitted by antenna 2. \square

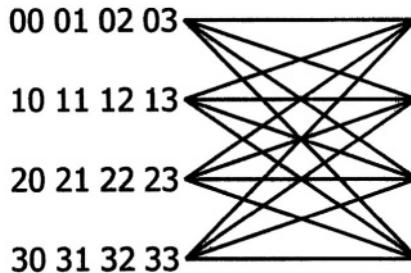


Figure 20.20. A space–time coding scheme with $t = 2$.

4.4 Space–time codes when CSI is not available

In a rapidly-changing mobile environment, or when long training sequences are not allowed, the assumption of perfect CSI at the receiver may not be valid. In the absence of CSI at the receiver, [HocMar00,MarHocHas00] advocate *unitary space–time modulation*, a technique which circumvents the use of training

symbols (which for maximum throughput should occupy half of the transmission interval, as seen before). Here the information is carried on the subspace that is spanned by orthonormal signals that are sent. This subspace survives multiplication by the unknown channel-gain matrix. A scheme based on *differential unitary space-time* signals is advocated in [HocSwe00]. High-rate constellations with excellent performance, obtained via algebraic techniques, are described in [HasHocShoSwe01].

5. Suboptimum receiver interfaces

From the capacity results described above we can see that extremely large spectral efficiencies can be achieved on a wireless link if the number of transmit and receive antennas is large. However, as t and r increase, the complexity of space-time coding with maximum-likelihood detection may become too large. This motivates the design of suboptimal receiver interfaces whose complexity is lower than ML, and yet perform close to what predicted by the theory. We describe here some of these schemes, categorized as *linear* and *nonlinear*.

5.1 Linear interfaces

A linear interface operates a linear transformation \mathbf{A} of the received signal, with $\mathbf{A} = \mathbf{A}(\mathbf{H})$ a $t \times r$ matrix chosen so as to allow a simplification of the metrics used in the Viterbi algorithm employed for soft decoding of \mathcal{X} .

Zero-forcing interface. This consists of choosing $\mathbf{A} = \mathbf{H}^+$, where the superscript $+$ denotes the Moore-Penrose pseudoinverse of a matrix [HorJoh85]. If we assume $r \geq t$, then $\mathbf{H}^\dagger \mathbf{H}$ is invertible with probability 1, and we have

$$\mathbf{H}^+ = (\mathbf{H}^\dagger \mathbf{H})^{-1} \mathbf{H}^\dagger \quad (20.68)$$

so that

$$\mathbf{H}^+ \mathbf{Y} = \mathbf{X} + \mathbf{H}^+ \mathbf{Z} \quad (20.69)$$

which shows that the spatial interference is completely removed from the received signal, while the noise is made nonwhite. The metric used with this interface is $\|\mathbf{H}^+ \mathbf{Y} - \mathbf{X}\|^2$. The conditional PEP is given by [BigTarTul02]

$$P(\mathbf{X} \rightarrow \hat{\mathbf{X}}) = \mathbb{E} \left[Q \left(\frac{\|\Delta\|^2}{\sqrt{2 N_0 \text{Tr}[(\mathbf{H}^\dagger \mathbf{H})^{-1} \Delta \Delta^\dagger]}} \right) \right] \quad (20.70)$$

where $\Delta \triangleq \mathbf{X} - \hat{\mathbf{X}}$.

Linear MMSE interface. Here we choose the matrix \mathbf{A} so as to minimize the mean-square value of the spatial interference plus noise. We

have

$$\mathbf{A} = (\mathbf{H}^\dagger \mathbf{H} + \delta_s \mathbf{I}_t)^{-1} \mathbf{H}^\dagger$$

where $\delta_s \triangleq N_0/E_s$, E_s the average energy per coded symbol. We have [Big-TarTul02]

$$P(\mathbf{X} \rightarrow \widehat{\mathbf{X}}) = \mathbb{E} \left[Q \left(\frac{\|\Delta\|^2 + 2((\mathbf{H}^\dagger \mathbf{H} + \delta_s \mathbf{I}_t)^{-1} \mathbf{H}^\dagger \mathbf{H} - \mathbf{I}_t) \mathbf{X}, \Delta)}{\sqrt{2N_0 \|\mathbf{H}(\mathbf{H}^\dagger \mathbf{H} + \delta_s \mathbf{I}_t)^{-1} \Delta\|^2}} \right) \right] \quad (20.71)$$

where $\Delta \triangleq \mathbf{X} - \widehat{\mathbf{X}}$. Notice that as $\delta_s \rightarrow 0$ the right-hand side of (20.71) tends to the PEP of the zero-forcing detector, as it should.

Performance of linear interfaces. Asymptotic analysis carried out for a large number of antennas [BigTarTul02] shows that these linear interfaces exhibit a PEP close to that of ML interface only for $r \gg t$; otherwise the performance loss may be substantial. As an example, Fig. 20.21 shows the error probability of a multiple-antenna system where the binary (8, 4, 4) Reed-Muller code is used by splitting its code words evenly between 2 transmit antennas. The word error probabilities shown are obtained through Monte Carlo simulation. Binary PSK is used, and the code rate is 1 bit per channel use. It is seen that for $r = 2$ both MMSE and ZF interface exhibit a considerable performance loss with respect to ML, while for $r = 8$ the losses are very moderate.

5.2 Nonlinear interfaces

The task of reducing the spatial interference affecting the received signal can be accomplished by first processing \mathbf{Y} linearly, then subtracting from the result an estimate of the spatial interference obtained from preliminary decisions $\widehat{\mathbf{X}}$ on the transmitted code word \mathbf{X} . The metric used for decoding is $\|\tilde{\mathbf{Y}} - \mathbf{X}\|$, where

$$\tilde{\mathbf{Y}} \triangleq \mathbf{G}\mathbf{Y} - \mathbf{L}\widehat{\mathbf{X}} \quad (20.72)$$

for a suitable choice of the two matrices \mathbf{G} and \mathbf{L} .

BLAST interface. One nonlinear interface is called BLAST (this stands for Bell Laboratories Layered Space-Time Architecture). Several implementations of the basic BLAST idea are possible, two of them being the *vertical MMSE-BLAST* interface and the *vertical ZF-BLAST* interface: these arise from the minimization of the mean-square error of the disturbance $\tilde{\mathbf{Y}} - \mathbf{X}$ with or without considering noise, respectively.

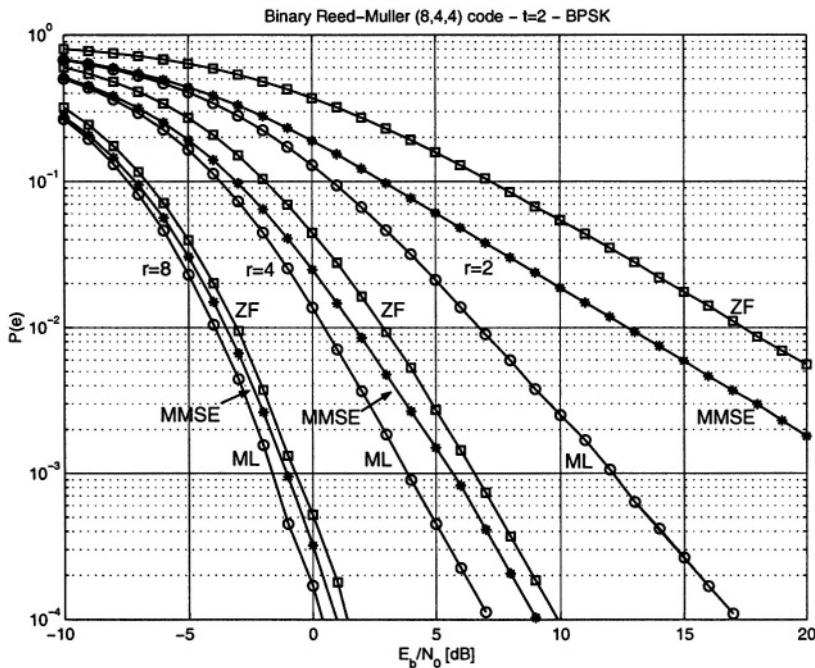


Figure 20.21. Performance of the binary (8, 4, 4) Reed-Muller code with binary PSK over a channel with $t = 2$ transmit antennas and r receive antennas with ML, MMSE, and ZF interfaces.

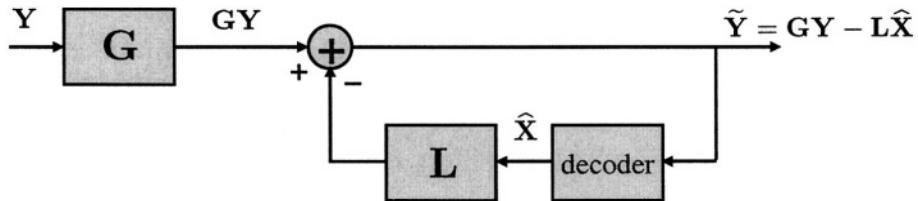


Figure 20.22. General structure of a nonlinear interface.

Zero-forcing BLAST. Here we have

$$\mathbf{G} = \text{diag}^{-1}(\mathbf{R})\mathbf{Q}^\dagger \quad (20.73)$$

$$\mathbf{L} = \text{diag}^{-1}(\mathbf{R})\mathbf{R} - \mathbf{I}_t \quad (20.74)$$

where \mathbf{Q} and \mathbf{R} come from the QR factorization $\mathbf{H} = \mathbf{QR}$ [HorJoh85]

where \mathbf{Q} and \mathbf{R} come from the QR factorization $\mathbf{H} = \mathbf{QR}$ [HorJoh85]

Here $\tilde{\mathbf{Y}}$ can be given the form

$$\tilde{\mathbf{Y}} = \underbrace{\mathbf{X}}_{\textcircled{1}} + \underbrace{[\text{diag}^{-1}(\mathbf{R})\mathbf{R} - \mathbf{I}_t]\Delta}_{\textcircled{2}} + \underbrace{\text{diag}^{-1}(\mathbf{R})\mathbf{Q}^\dagger \mathbf{Z}}_{\textcircled{3}} \quad (20.75)$$

The three terms in expression above can be recognized as: ① useful term (free from spatial interference: this justifies the name “zero-forcing”); ② interference due to past wrong decisions; and ③ noise. Notice that the entries of the noise matrix are independent but *not* identically distributed (i.e., the noise is *colored*).

MMSE-BLAST. Here

$$\begin{cases} \mathbf{G} = \text{diag}^{-1}(\mathbf{S})\mathbf{S}^{-\dagger}\mathbf{H}^\dagger \\ \mathbf{L} = \text{diag}^{-1}(\mathbf{S})\mathbf{S} - \mathbf{I}_t \end{cases} \quad (20.76)$$

where \mathbf{S} is an upper triangular matrix derived from the Cholesky factorization [HorJoh85] $\mathbf{H}^\dagger \mathbf{H} + \delta_s \mathbf{I}_t = \mathbf{S}^\dagger \mathbf{S}$, with $\delta_s \triangleq N_0/E_s$ as before. As a result, the soft estimate $\tilde{\mathbf{Y}}$ can be written as

$$\begin{aligned} \tilde{\mathbf{Y}} &= \text{diag}^{-1}(\mathbf{S})\mathbf{S}^{-\dagger}\mathbf{H}^\dagger \mathbf{Y} - (\text{diag}^{-1}(\mathbf{S})\mathbf{S} - \mathbf{I}_t)\hat{\mathbf{X}} \\ &= \underbrace{(\mathbf{I}_t - \text{diag}^{-1}(\mathbf{S})\mathbf{S}^{-\dagger})\mathbf{X}}_{\textcircled{1}} + \underbrace{(\text{diag}^{-1}(\mathbf{S})\mathbf{S} - \mathbf{I}_t)\Delta}_{\textcircled{2}} \\ &\quad + \underbrace{\text{diag}^{-1}(\mathbf{S})\mathbf{S}^{-\dagger}\mathbf{H}^\dagger \mathbf{Z}}_{\textcircled{3}} \end{aligned} \quad (20.77)$$

where the three terms in the last expression are: ① the (biased) useful term; ② the interference due to past wrong decisions; and ③ colored noise.

Appendix: derivation of (20.19)

Using results from [Sil95] we obtain the asymptotic value of the mean capacity as

$$\frac{C}{t} \rightarrow C(\alpha, \rho) \triangleq \int_a^b \log(1 + \alpha^{-1} \rho x) \frac{\sqrt{(x-a)(b-x)}}{2\pi\alpha x} dx \quad (20.78)$$

Then, setting $x = 1 + \alpha + 2\sqrt{\alpha} \cos \theta$, we obtain

$$C(\alpha, \rho) = \frac{1}{\pi} \int_0^{2\pi} \log \left(1 + \alpha^{-1} \rho (1 + \alpha + 2\sqrt{\alpha} \cos \theta) \right) \frac{\sin^2 \theta}{1 + \alpha + 2\sqrt{\alpha} \cos \theta} d\theta \quad (20.79)$$

Now, we factor the logarithm argument as a polynomial in $e^{j\theta}$ and obtain:

$$1 + \alpha^{-1} \rho (1 + \alpha + 2\sqrt{\alpha} \cos \theta) = \alpha^{-1/2} \rho e^{-j\theta} (e^{j\theta} + \alpha^{1/2} w_+) (e^{j\theta} + \alpha^{1/2} w_-)$$

where we defined $w \triangleq 1 + \alpha^{-1} + \rho^{-1}$ and $w_{\pm} \triangleq (w \pm \sqrt{w^2 - 4/\alpha})/2$. From the inequality

$$(\alpha^{1/2} w_+ - 1)(\alpha^{1/2} w_- - 1) = -\alpha^{-1/2}(1 - \alpha^{1/2})^2 - \alpha^{1/2} \rho^{-1} < 0$$

we have that $\alpha^{1/2} w_- < 1$ and $\alpha^{1/2} w_+ > 1$ so that we can write the logarithm argument as

$$1 + \alpha^{-1} \rho (1 + \alpha + 2\sqrt{\alpha} \cos \theta) = w_+ \rho \left(1 + \frac{\exp(j\theta)}{\alpha^{1/2} w_+} \right) \left(1 + \frac{\alpha^{1/2} w_-}{\exp(j\theta)} \right)$$

Thus, we obtain the following series expansion

$$\begin{aligned} & \ln(1 + \alpha^{-1} \rho (1 + \alpha + 2\sqrt{\alpha} \cos \theta)) \\ &= \ln(w_+ \rho) + \sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k} \left[\left(\frac{\exp(j\theta)}{\alpha^{1/2} w_+} \right)^k + \left(\frac{\alpha^{1/2} w_-}{\exp(j\theta)} \right)^k \right] \end{aligned} \quad (20.80)$$

which is uniformly convergent in θ and can be integrated term-by-term. Now, we need to calculate

$$I_k \triangleq \frac{1}{\pi} \int_0^{2\pi} e^{jk\theta} \frac{\sin^2 \theta}{1 + \alpha + 2\sqrt{\alpha} \cos \theta} d\theta \quad (20.81)$$

for every integer k . After some algebra, we obtain

$$I_k = \begin{cases} 1 & k = 0 \\ -\frac{1}{2}\sqrt{\alpha} & |k| = 1 \\ (-1)^{k+1} \frac{1-\alpha}{2\alpha} \alpha^{|k|/2} & |k| > 1 \end{cases} \quad (20.82)$$

The termwise integration of the series yields

$$\begin{aligned}
 C(\alpha, \rho) &= \ln(w_+ + \rho) - \frac{1}{2\alpha} \left[w_+^{-1} + \alpha w_- \right] + \frac{1-\alpha}{2\alpha} \sum_{k=1}^{\infty} \frac{1}{k} \left[w_+^{-k} + (\alpha w_-)^k \right] \\
 &= \ln(w_+ + \rho) - w_- - \frac{1-\alpha}{2\alpha} \left[\ln(1 - w_+^{-1}) + \ln(1 - \alpha w_-) \right] \\
 &= \ln(w_+ + \rho) - w_- - \frac{1-\alpha}{\alpha} \ln(1 - w_+^{-1}) \\
 &= \frac{1}{\alpha} \ln(w_+ + \rho) - w_- - \frac{1-\alpha}{\alpha} \ln((w_+ - 1)\rho) \\
 &= \frac{1}{\alpha} \ln(w_+ + \rho) - w_- + \frac{1-\alpha}{\alpha} \ln(1 - w_-)
 \end{aligned} \tag{20.83}$$

Both series expansions in brackets converge because $w_+ > 1$ and $\alpha w_- < 1$ since

$$(w_+ - 1)(\alpha w_- - 1) = 2 - w_+ - \alpha w_- = -\frac{(w_+ - 1)^2}{w_+} < 0 \tag{20.84}$$

Notations and Acronyms

- ☞ a.s., almost surely
- ☞ AWGN, Additive white Gaussian noise
- ☞ CSI, Channel state information
- ☞ iid, Independent and identically distributed
- ☞ LHS, Left-hand side
- ☞ ln, Natural logarithm
- ☞ log, Logarithm in base 2
- ☞ ML, Maximum-likelihood
- ☞ pdf, Probability density function
- ☞ RHS, Right-hand side
- ☞ RV, Random variable
- ☞ SNR, Signal-to-noise ratio
- ☞ TCM, Trellis-coded modulation
- ☞ a^* , Conjugate of the complex number a
- ☞ $(a)_+ \triangleq \max(0, a)$, Equal to a if $a > 0$, equal to 0 otherwise.
- ☞ \mathbf{A}^\dagger , Conjugate (or Hermitian) transpose of matrix \mathbf{A}
- ☞ $\|\mathbf{A}\| \triangleq \sqrt{\sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2}$, Frobenius norm of the $m \times n$ matrix $\mathbf{A} = (a_{ij})$, $i = 1, \dots, m$, $j = 1, \dots, n$
- ☞ \mathbb{C} , The set of complex numbers
- ☞ $\mathbb{E}[X]$, Expectation of the RV X
- ☞ \mathbf{I}_n , The $n \times n$ identity matrix
- ☞ \Im , Imaginary part
- ☞ $Q(x) \triangleq (2\pi)^{-1/2} \int_x^\infty \exp(-z^2/2) dz$, The Gaussian tail function
- ☞ \Re , Real part
- ☞ \mathbb{R} , The set of real numbers
- ☞ \triangleq , Equal by definition
- ☞ $X \sim \mathcal{N}(\mu, \sigma^2)$, X is a real Gaussian RV with mean μ and variance σ^2
- ☞ $X \sim \mathcal{N}_c(\mu, \sigma^2)$, X is a circularly-distributed complex Gaussian RV with mean μ and $\mathbb{E}[|X|^2] = \sigma^2$

References

- [BaiSil02] Z. D. Bai and J. W. Silverstein, “CLT for linear spectral statistics of large dimensional sample covariance matrices,” *Annals of Probability*, to be published, 2003.
- [BenBig99] S. Benedetto and E. Biglieri, *Principles of Digital Transmission with Wireless Applications*. New York: Kluwer/Plenum, 1999.
- [BigTarTul02] E. Biglieri, G. Taricco, and A. Tulino, “Performance of space–time codes for a large number of antennas,” *IEEE Trans. Inform. Theory*, Vol. 48, No. 7, pp. 1794–1803, Jul. 2002.
- [BigCaiTar01] E. Biglieri, G. Caire, and G. Taricco, “Limiting performance of block-fading channels with multiple antennas,” *IEEE Trans. Inform. Theory*, Vol. 47, No. 4, pp. 1273–1289, May 2001.
- [BigProSha98] E. Biglieri, J. Proakis, and S. Shamai (Shitz), “Fading channels: Information-theoretic and communication aspects,” *IEEE Trans. Inform. Theory*, Vol. 44, No. 6, 50th Anniversary Issue, pp. 2619–2692, October 1998.
- [CaiKnoHum98] G. Caire, R. Knopp and P. Humbert, “System capacity of F-TDMA cellular systems,” *IEEE Trans. Commun.*, Vol. 46, No. 12, pp. 1649–1661, Dec. 1998.
- [CaiTarBig98] G. Caire, G. Taricco and E. Biglieri, “Optimal power control for minimum outage rate in wireless communications,” *Proc. of IEEE ICC ’98*, Atlanta, GA, June 1998.
- [CaiTarBig99] G. Caire, G. Taricco and E. Biglieri, “Optimal power control for the fading channel,” *IEEE Trans. Inform. Theory*, Vol. 45, No. 5, pp. 1468–1489, July 1999.
- [ChaHwaLin97] J.-J. Chang, D.-J. Hwang, and M.-C. Lin, “Some extended results on the search for good convolutional codes,” *IEEE Trans. Inform. Theory*, Vol. 43, No. 5, pp. 1682–1697, Sept. 1997.
- [ChiRimTel01] N. Chiurtu, B. Rimoldi, and E. Telatar, “Dense multiple antenna systems,” *Proc. of IEEE ITW 2001*, Cairns, Australia, pp. 108–109, Sept. 2–7, 2001.
- [ChuTseKah98] C.-N. Chuah, D. Tse, and J. M. Kahn, “Capacity of multi-antenna array systems in indoor wireless environment,” *Proc. of IEEE GLOBECOM ’98*, Sydney, Australia, Nov. 8–12, 1998.

- [CovTho91] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
- [Foschini96] G. J. Foschini, “Layered space–time architecture for wireless communication in a fading environment when using multi-element antennas,” *Bell Labs Tech. J.*, Vol. 1, No. 2, pp. 41–59, Autumn 1996.
- [FosGan98] G. J. Foschini and M. J. Gans, “On limits of wireless communications in a fading environment when using multiple antennas,” *Wireless Personal Communications*, Vol. 6, No. 3, pp. 311–335, March 1998.
- [Gal68] R. G. Gallager, *Information Theory and Reliable Communication*. New York: J. Wiley & Sons, 1968.
- [GueFitBelKuo99] J.-C. Guey, M. P. Fitz, M. R. Bell, and W.-Y. Kuo, “Signal design for transmitter diversity wireless communication systems over Rayleigh fading channels,” *IEEE Trans. Commun.*, Vol. 47, No. 4, pp. 527–537, April 1999.
- [HasHocShoSwe01] B. Hassibi, B. M. Hochwald, A. Shokrollahi, and W. Sweldens, “Representation theory for high-rate multiple-antenna code design,” *IEEE Trans. Inform. Theory*, Vol. 47, No. 6, pp. 2335–2367, Sept. 2001.
- [HocMar98] B. Hochwald and T. Marzetta, “Space–time modulation scheme for unknown Rayleigh fading environment,” *Proc. of 36th Annual Allerton Conference on Communication, Control and Computing*, Allerton House, Monticello, IL, September 1998.
- [HocMar00] B. Hochwald and T. Marzetta, “Unitary space–time modulation for multiple-antenna communication in Rayleigh flat-fading,” *IEEE Trans. Inform. Theory*, Vol. 46, No. 2, pp. 543–564, March 2000.
- [HocMarTar02] B. M. Hochwald, T. L. Marzetta, and V. Tarokh, “Multi-antenna channel-hardening and its implications for rate feedback and scheduling,” *IEEE Trans. Inform. Theory*, submitted for publication, May 2002.
- [HocSwe00] B. Hochwald and W. Sweldens, “Differential unitary space time modulation,” *IEEE Trans. Commun.*, Vol. 48, No. 12, pp. 2041–2052, Dec. 2000.
- [HorJoh85] R. Horn and C. Johnson, *Matrix Analysis*. New York: Cambridge University Press, 1985.

- [Knopp97] R. Knopp, *Coding and Multiple-Accessing over Fading Channels*, PhD dissertation, EPFL, Lausanne (Switzerland), and Institut Eurécom, Sophia Antipolis (France), 1997.
- [KnoHum00] R. Knopp and P. A. Humblet, “On coding for block-fading channels,” *IEEE Trans. Inform. Theory*, Vol. 46, No. 1, pp. 189–205, Jan. 2000.
- [MaGian02] X. Ma and G. B. Giannakis, “Full-diversity full-rate complex-field space–time coding,” *IEEE Trans. Sig. Proc.*, Vol. 51, No. 11, pp. 2917–2930, Nov. 2003.
- [MalLei97] E. Malkamäki and H. Leib, “Coded diversity schemes on block fading Rayleigh channels,” *Proc. of IEEE ICUPC ’97*, San Diego, CA, Oct. 1997.
- [MalLei99] E. Malkamäki and H. Leib, “Coded diversity on block-fading channels,” *IEEE Trans. Inform. Theory*, Vol. 45, No. 2, pp. 771–781, March 1999.
- [Marzetta99] T. L. Marzetta; “BLAST training: Estimating channel characteristics for high capacity space–time wireless,” *Proc. of 37th Annual Allerton Conference on Communication, Control and Computing*, Allerton House, Monticello, IL, pp. 958–966. Sept. 22–24, 1999.
- [MarHoc99] T. L. Marzetta and B. M. Hochwald, “Capacity of a mobile multiple-antenna communication link in Rayleigh flat fading,” *IEEE Trans. Inform. Theory*, Vol. 45, No. 1, pp. 139–157, January 1999.
- [MarHocHas00] T. L. Marzetta, B. M. Hochwald, and B. Hassibi, “New approach to single-user multiple-antenna wireless communication,” *Proc. of CISS 2000*, Princeton University, pp. WA4-16–WA4-21, March 15–17, 2000.
- [MouSimSen03] A.L. Moustakas, S.H. Simon, A.M. Sengupta, “MIMO Capacity Through Correlated Channels in the Presence of Correlated Interferers and Noise: A (Not So) Large N Analysis,” *IEEE Trans. Inform. Theory* Vol. 49, No. 10, pp. 2545–2561, Oct. 2003.
- [OzaShaWyn94] L. Ozarow, S. Shamai, and A. D. Wyner, “Information theoretic considerations for cellular mobile radio,” *IEEE Trans. Vehic. Tech.*, Vol. 43, No. 2, pp. 359–378, May 1994.

- [RalCio98] G. Raleigh and J. Cioffi, "Spatio-temporal coding for wireless communication," *IEEE Trans. Commun.*, Vol. 46, No. 3, pp. 357–366, March 1998.
- [SenMit02] A. M. Sengupta and P. P. Mitra, "Capacity of multivariate channels with multiplicative noise: Random matrix techniques and large- N expansions for full transfer matrices," *submitted for publication*, 2002.
- [ShiKah99] D. Shiu and J. M. Kahn, "Design of high-throughput codes for multiple-antenna wireless systems," Submitted to *IEEE Trans. Inform. Theory*, January 1999.
- [ShiKah99a] D. Shiu and J. M. Kahn, "Layered space-time codes for wireless communications using multiple transmit antennas," *Proc. of IEEE ICC'99*, Vancouver, BC, June 6–10, 1999.
- [Sil95] J. W. Silverstein, "Strong convergence of the empirical distribution of eigenvalues of large dimensional random matrices," *Journal of Multivariate Analysis*, Vol. 55, pp. 331–339, 1995.
- [SmiSha02] P. J. Smith and M. Shafi, "On a Gaussian approximation to the capacity of wireless MIMO systems," *Proc. of IEEE ICC 2002*, pp. 406–410, New York, April 28–May 2, 2002.
- [SteDum99] A. S. Stefanov and T. M. Duman, "Turbo coded modulation for systems with transmit and receive antenna diversity," *Proc. of IEEE GLOBECOM '99*, Rio de Janeiro, Brazil, pp. 2336–2340, Dec. 5–9, 1999.
- [Szego39] G. Szegö, *Orthogonal Polynomials*. Americal Mathematical Society, Providence, RI, 1939.
- [TarBig03] G. Taricco and E. Biglieri, "Space-time decoding with imperfect channel estimation," *IEEE Trans. Wireless Communications*, submitted for publication, August 2003.
- [TarLo98] V. Tarokh and T. K. Y. Lo, "Principal ratio combining for fixed wireless applications when transmitter diversity is employed," *IEEE Commun. Letters*, Vol. 2, No. 8, pp. 223–225, August 1998.
- [TarSesCal98] V. Tarokh, N. Seshadri, and A. R. Calderbank, "Space-time codes for high data rate wireless communication: Performance criterion and code construction," *IEEE*

- Trans. Inform. Theory* Vol. 44, No. 2, pp. 744–765, March 1998.
- [TarJafCal99] V. Tarokh, H. Jafarkhani, and A. R. Calderbank, “Space–time block codes from orthogonal designs,” *IEEE Trans. Inform. Theory*, Vol. 45, No. 5, pp. 1456–1467, July 1999.
- [Telatar99] E. Telatar, “Capacity of multi-antenna Gaussian channels,” *European Trans. Telecomm.*, Vol. 10, No. 6, pp. 585–595, November–December 1999.
- [TseHanly98] D. Tse and V. Hanly, “Multi-access fading channels—Part I: Polymatroid structure, optimal resource allocation and throughput capacities,” *IEEE Trans. Inform. Theory*, Vol. 44, No. 7, pp. 2796–2815, November 1998.
- [WanGia02] Z. Wang and G. B. Giannakis, “Outage mutual information of space–time MIMO channels,” *Fortieth Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, October 2–4, 2002.
- [ZheTse02] L. Zheng and D. N. C. Tse, “Communication on the Grassman manifold: A geometric approach to the non-coherent multiple-antenna channel,” *IEEE Trans. Inform. Theory*, Vol. 48, No. 2, pp. 359–383, February 2002.
- [ZheTse03] L. Zheng and D. N. C. Tse, “Diversity and multiplexing: A fundamental tradeoff in multiple antenna channels,” *IEEE Trans. Inform. Theory*, Vol. 49, No. 5, pp. 1073–1096, May 2003.

Chapter 21

DIAGONAL STBC'S FOR FADING ISI CHANNELS: CODE DESIGN AND EQUALIZATION

Robert Schober

*Department of Electrical and Computer Engineering
University of British Columbia*

rschober@ece.ubc.ca

Wolfgang H. Gerstacker

*Institute for Mobile Communication
University of Erlangen–Nuernberg*
gersta@LNT.de

Lutz H.-J. Lampe

*Department of Electrical and Computer Engineering
University of British Columbia*
llampe@ece.ubc.ca

Subbarayan Pasupathy

*Department of Electrical and Computer Engineering
University of Toronto*
pas@comm.utoronto.ca

Abstract

In this chapter, we design and optimize matrix–symbol–based space–time block codes (STBC's) for transmission over fading intersymbol interference (ISI) channels. We show that STBC's employing diagonal code matrices exclusively facilitate the successful application of suboptimum equalization techniques for the practically important case when only a single receive antenna is available. Three different types of diagonal STBC's are optimized for fading ISI channels and their performances are compared for decision–feedback equalization (DFE) and decision–feedback sequence estimation (DFSE), respectively. The robustness of the designed codes against variations of the characteristics of the fading ISI

channel and the dependence of the equalizer performance on the STBC structure are investigated.

It is shown that for data rates of $R \leq 2$ bits/(channel use) and typical multipath fading channels diagonal STBC's outperform Alamouti's code if suboptimum equalization schemes are adopted.

Keywords: Space–time block codes, multiple antennas, frequency–selective fading channels, equalization.

21.1 Introduction

It is well known that space–time (ST) codes can significantly improve the performance of wireless communication systems, e.g. [1–4]. Initially, the investigation of ST coding was limited to flat fading channels. However, since most wireless channels are impaired by intersymbol interference (ISI), more recently considerable research effort has been dedicated to ST coding for frequency–selective channels, cf. [5–14].

A fundamental problem of ST coding for fading ISI channels is to find ST codes which achieve a high performance and at the same time allow the application of suboptimum low–complexity equalization techniques such as decision–feedback equalization (DFE) or reduced–state sequence estimation (RSSE) [15, 16]. This is especially true if multiple transmit antennas along with only a single receive antenna are applied. In this chapter, we will focus on this important scenario which is likely to occur in down–link transmission (base station to mobile station).

For ST trellis codes [3], reduced–state decoding with high performance is only possible if the cascade of ST encoder and ISI channel can be interpreted as a delay diversity scheme [17–19]. Recently, several ST block coding schemes based on processing of entire bursts have been proposed, e.g. [5, 7, 8, 10, 11, 14]. The scheme in [8] is a generalization of Alamouti's ST block code (STBC) [2], whereas the schemes in [5, 7, 11] and [10, 14] are based on orthogonal frequency division multiplexing (OFDM) and on single–carrier transmission combined with frequency domain equalization, respectively. These burst–based STBC's achieve very good performance for block fading channels with constant or very slowly varying impulse response inside each burst and, in addition, in most cases low–complexity equalization methods exist, cf. [21] for a recent overview and comparison of several schemes. However, it is not clear how these schemes can be extended to channels with time–variant behavior. Therefore, in this contribution we adopt a different approach and investigate the applicability of (matrix–) symbol–based STBC's for transmission over fading ISI channels, since in this case, an extension to time–variant channels can be easily accomplished by adapting the equalizer filters using least–mean–square (LMS) or recursive least–squares (RLS) algorithms.

The only matrix–symbol–based STBC that has been considered for fading ISI channels so far is Alamouti's code [2]. It has been found that with simple conventional equalizers a good performance can only be achieved if at least two receive antennas are available, e.g. [6, 9, 13]. More recently, however, suboptimum symbol–by–symbol [22] and trellis–based [23] equalizers, which are based on widely linear (WL) signal processing [24], have been proposed, see also [12]. These WL equalizers can achieve a high performance with a single receive antenna and will serve as a benchmark for the scheme proposed here.

In this chapter, we *optimize* ST block codes for fading ISI channels under the constraint that suboptimum equalization is used at the receiver. This naturally leads to *diagonal* STBC's where only one transmit antenna is active at any time. For diagonal STBC's employing N_T transmit antennas and one receive antenna an equivalent multiple–input multiple–output (MIMO) system with N_T inputs and N_T outputs exists. This equivalent MIMO system facilitates the application of both DFE and efficient RSSE techniques. Diagonal STBC's were reported first by DaSilva and Sousa [25]¹. Later, Hughes [26] and Hochwald and Sweldens [27] proposed diagonal STBC's which enable differential encoding. For optimization and comparison of different diagonal STBC's, we adapt the general design criterion for STBC's for fading ISI channels introduced in [28, 29] to the problem at hand.

Our simulation results show that for data rates of $R \leq 2$ bits/(channel use) diagonal STBC's combined with suboptimum equalization allow significant performance improvements over conventional (single antenna) transmission at the expense of a moderate increase in receiver complexity and compare favorably with the scheme proposed in [12].

This chapter is organized as follows. In Section 21.2, the transmission model is introduced, the adopted design criterion for STBC's is briefly reviewed, and the reasons for the application of diagonal STBC's are explained in detail. Various different diagonal STBC's are optimized and compared in Section 21.3. In Section 21.4, an equivalent MIMO model for diagonal STBC's is established and suboptimum equalizers are discussed. Simulation results are presented in Section 21.5 and some conclusions are drawn in Section 21.6.

21.2 Preliminaries

21.2.1 Notation

Bold upper case (X) and lower case (\mathbf{x}) letters denote matrices and vectors, respectively, $\det(\cdot)$, $\text{tr}(\cdot)$, $(\cdot)^T$, $(\cdot)^H$, and $(\cdot)^*$ refer to the determinant and the

¹DaSilva and Sousa refer to their transmit diversity scheme as “fading–resistant modulation” [25]. Nevertheless, this method can be interpreted as a diagonal STBC.

trace of a matrix, transposition, Hermitian transposition, and complex conjugation, respectively. \mathbf{I}_X and $\mathbf{0}_X$ are the $X \times X$ identity matrix and the all-zero column vector of size X , respectively. $\mathcal{E}\{\cdot\}$, mod, $j \triangleq \sqrt{-1}$, and $\lfloor X \rfloor$ denote expectation, the modulo operator, the imaginary unit, and the largest integer smaller than or equal to X , respectively. Throughout this chapter, all signals are represented by their complex-baseband equivalents.

21.2.2 Transmission Model

In this chapter, we focus on ST block coded transmission employing N_T , $N_T \geq 2$, transmit and $N_R = 1$ receive antennas. The STBC matrix in its most general form is given by

$$\mathbf{C}[k_N] = \begin{pmatrix} c_0[Nk_N] & \cdots & c_{N_T-1}[Nk_N] \\ \vdots & \ddots & \vdots \\ c_0[Nk_N + N - 1] & \cdots & c_{N_T-1}[Nk_N + N - 1] \end{pmatrix} \quad (2.1)$$

with matrix discrete-time index $k_N \triangleq \lfloor k/N \rfloor$ ($k \in \mathbb{Z}$: symbol discrete-time index), i.e., at time $k = Nk_N + \kappa$, $0 \leq \kappa \leq N - 1$, the symbol $c_\nu[k]$ is transmitted from antenna ν , $0 \leq \nu \leq N_T - 1$. Thereby, $\mathbf{C}[\cdot]$ is taken from a set \mathcal{C} of 2^{N_b} code matrices \mathbf{C}^α , $0 \leq \alpha \leq 2^{N_b} - 1$. $c_\nu^\alpha[\kappa]$, $0 \leq \nu \leq N_T - 1$, $0 \leq \kappa \leq N - 1$, denotes the element of \mathbf{C}^α in column ν and row κ . $N_b = RN$ information bits are mapped to the elements of \mathcal{C} , where R is the data rate in bits per channel use. We adopt the normalization $\sum_{\nu=0}^{N_T-1} \mathcal{E}\{|c_\nu[\kappa]|^2\} = 1$ in order to make the transmitted energy independent of the number of transmit antennas N_T .

The ST block coded signal is transmitted over a fading ISI channel and after receive filtering and T -spaced sampling (T denotes the symbol interval) the received signal $r[k]$ is modeled as

$$r[k] = \sum_{\nu=0}^{N_T-1} \sum_{l=0}^{L-1} h_\nu[l] c_\nu[k-l] + n[k], \quad (2.2)$$

where $h_\nu[l]$ and $n[k]$ refer to the l th coefficient of the discrete-time overall channel impulse response (CIR) between the ν th transmit antenna and the receive antenna and discrete-time additive white Gaussian noise (AWGN), respectively. The discrete-time CIR's $h_\nu[\cdot]$, $0 \leq \nu \leq N_T - 1$, are truncated to the same length L , where $\sum_{l=-\infty}^{-1} \mathcal{E}\{|h_\nu[l]|^2\} + \sum_{l=L}^{\infty} \mathcal{E}\{|h_\nu[l]|^2\} \approx 0$ is assumed. Furthermore, $h_\nu[\cdot]$ includes the combined effects of transmit filter, multipath Rayleigh fading channel, and (square-root Nyquist) receiver input filter. Therefore, for a given ν the discrete-time CIR coefficients $h_\nu[l]$ are mutually correlated zero-mean Gaussian random variables. For simplicity,

for different transmit antennas the CIR coefficients are assumed to be mutually uncorrelated, i.e., $\mathcal{E}\{h_{\nu_1}[l_1]h_{\nu_2}^*[l_2]\} = 0$, $\nu_1 \neq \nu_2$, $\forall l_1, l_2$, but to have identical statistical properties. Nevertheless, most concepts introduced in this chapter can be extended in a straightforward way to mutually correlated CIR's. We assume that $h_\nu[\cdot]$ is constant over one burst but varies from burst to burst, e.g. due to frequency hopping (block fading model). Such a model is realistic for moderate burst lengths and moderate fading velocities. As customary in practice, we adopt a (suboptimum) fixed square-root Nyquist receiver input filter instead of the optimum matched filter. This facilitates implementation and usually causes only a negligible loss in performance [30].

Because of an appropriate normalization, $\sum_{l=0}^{L-1} \mathcal{E}\{|h_\nu[l]|^2\} = 1$, $0 \leq \nu \leq N_T - 1$, is valid and the noise variance is given by $\sigma_n^2 \stackrel{\Delta}{=} \mathcal{E}\{|n[k]|^2\} = N_0/(RE_b)$, where N_0 and E_b denote the single-sided power spectral density of the underlying continuous-time passband noise process and the mean received energy per bit, respectively.

21.2.3 Design Criterion for STBC's

In [28, 29], a lower bound for the pairwise error probability (PEP) $P_e(\alpha, \beta)$, i.e., the probability that \mathbf{C}_α is transmitted and \mathbf{C}_β , $\beta \neq \alpha$, is detected, is given for the transmission scheme described in the previous subsection. This bound is related to the classical matched filter bound, e.g. [31], since it was obtained by considering the transmission of a single STBC matrix. A Chernoff bound on this PEP is given by [29]

$$P(\alpha, \beta) \leq \frac{1}{\prod_{q=1}^Q \lambda_q \left(\frac{1}{4\sigma_n^2} \right)^Q}, \quad (2.3)$$

where Q and λ_q , $1 \leq q \leq Q$, are the rank and the non-zero eigenvalues of matrix $\mathbf{C}_m^{\alpha|\beta} \Phi_{hh} (\mathbf{C}_m^{\alpha|\beta})^H$, respectively. Thereby, Φ_{hh} is the autocorrelation matrix (ACM) of all CIR coefficients and is defined as

$$\Phi_{hh} \stackrel{\Delta}{=} \mathcal{E}\{\mathbf{h}\mathbf{h}^H\} \quad (2.4)$$

with

$$\mathbf{h} \stackrel{\Delta}{=} [\mathbf{h}_0^T \dots \mathbf{h}_{N_T-1}^T]^T \quad (2.5)$$

$$\mathbf{h}_\nu \stackrel{\Delta}{=} [h_\nu[0] \dots h_\nu[L-1]]^T. \quad (2.6)$$

Matrix

$$\mathbf{C}_m^{\alpha|\beta} \stackrel{\Delta}{=} \mathbf{C}_m^\beta - \mathbf{C}_m^\alpha \quad (2.7)$$

denotes the difference of two modified STBC matrices \mathbf{C}_m^x , $x \in \{\alpha, \beta\}$, which are uniquely specified by the original STBC matrices \mathbf{C}^x via

$$\mathbf{C}_m^x \triangleq [\mathbf{C}_0^{x'} \dots \mathbf{C}_{N_T-1}^x], \quad (2.8)$$

$$\mathbf{C}_\nu^x \triangleq [\mathbf{c}_{1\nu}^x \dots \mathbf{c}_{L\nu}^x], \quad (2.9)$$

$$\mathbf{c}_{l\nu}^x \triangleq [\mathbf{0}_{l-1}^T \mathbf{c}_\nu^x[0] \dots \mathbf{c}_\nu^x[N-1] \mathbf{0}_{L-l}^T]^T. \quad (2.10)$$

From Eq. (2.3) it has been deduced in [28, 29] that the STBC set \mathcal{C} should be selected to maximize the diversity order Q . For most practical channels it can be assumed that a STBC can be found for which matrix $\mathbf{C}_m^{\alpha|\beta} \Phi_{hh} (\mathbf{C}_m^{\alpha|\beta})^H$ has full rank for all pairs (α, β) , $\alpha \neq \beta$, provided that the CIR truncation length L and N (see next subsection) are chosen appropriately. Therefore, we may define the distance measure

$$d^2(\alpha, \beta) \triangleq \det \left(\mathbf{C}_m^{\alpha|\beta} \Phi_{hh} (\mathbf{C}_m^{\alpha|\beta})^H \right). \quad (2.11)$$

If $\mathbf{C}_m^{\alpha|\beta} \Phi_{hh} (\mathbf{C}_m^{\alpha|\beta})^H$ is not full rank, in Eq. (2.11) the determinant of the matrix has to be replaced by the product of the Q non-zero eigenvalues.

Now, a reasonable STBC design criterion for fading ISI channels is to maximize the minimum distance²

$$d_{min}^2 \triangleq \operatorname{argmin}_{\substack{(\alpha, \beta) \\ \beta \neq \alpha}} \left\{ d^2(\alpha, \beta) \right\}. \quad (2.12)$$

The optimum code among all STBC's which achieve the maximum possible diversity order is that one which maximizes d_{min}^2 . Note that for $L = 1$ and $N = N_T$ the above design rule is identical to the conventional design rule for flat fading channels [3].

21.2.4 Why Diagonal STBC's?

A primary goal of this chapter is to find STBC's which also provide a performance improvement compared to single-antenna transmission if suboptimum (low-complexity) equalization strategies are applied. The straightforward approach to this problem would be to optimize the STBC given the suboptimum equalization scheme (e.g. DFE) used. For this purpose we would need to derive a corresponding design criterion. Unfortunately, this is a very difficult if not impossible task and a practical design criterion is not likely to exist. On the other hand, the unconstrained maximization of the minimum distance

²A related design criterion for ST trellis codes has been proposed independently by Liu et al. [32].

d_{\min}^2 will lead to STBC's which perform very well for computationally complex maximum-likelihood sequence detection (MLSD) [28, 29] but may be not suitable at all for application of low-complexity suboptimum equalization techniques. Therefore, we have to impose additional constraints on the structure of the STBC matrix to get viable solutions.

First, we should be aware that, in general, the $N_T N$ elements of the transmitted STBC matrix $\mathbf{C}[\cdot]$ are not correlated. If we consider e.g. Alamouti's code with $N = N_T = 2$,

$$\mathbf{C}[k_2] = \frac{1}{\sqrt{2}} \begin{pmatrix} b_0[k_2] & b_1[k_2] \\ -b_1^*[k_2] & b_0^*[k_2] \end{pmatrix} \quad (2.13)$$

is valid, where $b_0[\cdot]$ and $b_1[\cdot]$ are independently taken from the same M -ary phase-shift keying (MPSK) or M -ary quadrature amplitude modulation (MQAM) signal alphabet [2]. Clearly, in this case, for rotationally invariant complex signal constellations such as 4PSK or 16QAM the elements of $\mathbf{C}[\cdot]$ are mutually uncorrelated. Since the four elements of $\mathbf{C}[\cdot]$ are transmitted in two time steps, a separation of these elements at the receiver using equalizers which rely on second-order statistics is only possible if $N_R \geq 2$ receive antennas are employed. This problem could be slightly relaxed of course if oversampling of the received continuous-time signal was applied, cf. e.g. [33]. However, since, in general, the polyphase components of the oversampled continuous-time overall CIR are strongly correlated, the resulting receivers would be very sensitive to channel noise. For the special case of Alamouti's code a solution to this problem is given in [12] using widely linear signal processing [24]. However, for a general STBC the $N_T N$ elements of matrix $\mathbf{C}[\cdot]$ are not separable at the receiver. To circumvent the above mentioned problems, we transmit in N time steps only N symbols since in this case the (non-zero) elements of $\mathbf{C}[\cdot]$ can be separated at the receiver.

It has been shown in [28] that for mutually uncorrelated CIR coefficients the maximum possible diversity order $Q = LN_T$ can only be achieved, if N is chosen as $N = (N_T - 1)L + 1$. However, our investigations have shown that if only one transmit antenna is active at any time, $N = N_T$ yields the best performance for bit error rates (BER's) of practical interest, while $N > N_T$ is only beneficial at very low BER's. Therefore, for simplicity of presentation, in the rest of this chapter we restrict our attention to the case $N = N_T$, i.e., to diagonal STBC matrices $\mathbf{C}[\cdot]$.

21.3 Design of Diagonal STBC's

In this section, three different types of diagonal STBC's are discussed and optimized for the widely accepted hilly terrain (HT) channel model [34] using the criterion given in Eq. (2.12). The obtained designs are then compared based on an approximation for the BER for MLSD [28].

For diagonal STBC's the only non-zero elements of the code matrix \mathbf{C}^α are $c_\nu^\alpha[\nu], 0 \leq \nu \leq N_T - 1$. Therefore, for the following, we introduce the simpler notation $c_\nu^\alpha \triangleq c_\nu^\alpha[\nu]$.

21.3.1 Repetition Codes (RC's)

The simplest diagonal STBC's are repetition codes which will serve as a benchmark for more sophisticated codes. For N_T transmit antennas and RC's of rate R , $c_\nu^\alpha = b^\alpha, 0 \leq \nu \leq N_T - 1$, is valid, where b^α are MPSK or MQAM symbols with $M = 2^{N_T R}$. If we take in Eq. (2.12) into account that all c_ν^α are identical, we obtain $d_{min}^2 = x_h \cdot d_{b,min}^2$, where x_h is a constant which depends only on the channel correlation matrix Φ_{hh} and $d_{b,min}^2$ is the minimum squared Euclidean distance of the adopted MPSK or MQAM signal constellation. Therefore, as one would intuitively expect, in order to maximize d_{min}^2 for given R and N_T a scalar signal constellation which has the maximum minimum squared Euclidean distance has to be selected. In addition, a Gray labeling of the scalar signal constellation constitutes also a Gray labeling for the RC.

21.3.2 Hochwald/Sweldens Codes (HS-C's) [27]

HS-C's have been introduced in [27] for differential ST transmission. In this work, we focus on coherent transmission, i.e., differential encoding is not applied. Thus, the transmitted symbols are given by

$$c_\nu^\alpha = \exp\left(j\frac{2\pi u_\nu \alpha}{2^{N_T R}}\right), \quad (3.1)$$

$0 \leq \nu \leq N_T - 1, 0 \leq \alpha \leq 2^{N_T R} - 1$. The parameters $u_\nu, 0 \leq u_\nu \leq 2^{N_T R} - 1$, are integers³ which are optimized in [27] for differential detection and flat Rayleigh fading channels. Here, we have to optimize the parameter vector $\mathbf{u} \triangleq [u_0 \dots u_{N_T - 1}]$ for coherent reception and Rayleigh fading ISI channels. We note that since we assume that the CIR coefficients pertaining to different transmit antennas are mutually uncorrelated, Φ_{hh} is a block diagonal matrix. Therefore, for HS-C's the distance $d^2(\alpha, \beta)$ defined in Eq. (2.11) depends only on the difference $(\alpha - \beta) \bmod(2^{N_T R})$. Hence, the optimum parameter vector \mathbf{u}_{opt} can be obtained from

$$\mathbf{u}_{opt} = \underset{0 \leq u_0, \dots, u_{N_T - 1} \leq 2^{N_T R} - 1}{\operatorname{argmax}} \left\{ \min_{\alpha=1, \dots, 2^{N_T R} - 1} \left\{ d^2(\alpha, 0) \right\} \right\}, \quad (3.2)$$

³Since we do not employ differential encoding, we could also allow non-integer values for the u_ν . However, this complicates the code search and is expected to yield only marginal performance improvements.

i.e., we have to search over $2^{N_T^2 R}$ possible parameter vectors \mathbf{u} . Using similar methods as in [27] the search space may be reduced to some extent. However, we do not expand on this topic since HS-C's are mainly beneficial for low data rates and a small number of transmit antennas implying a moderate code search complexity.

21.3.3 Rotated MASK (R-MASK) and MQAM (R-MQAM) Constellations

For the flat fading channel, diagonal STBC's have been designed in [25] by rotating MASK constellations⁴. For this, we define the vector $\mathbf{b}^\alpha \triangleq [b_0^\alpha \dots b_{N_T-1}^\alpha]^T$ of MASK symbols and the rotated vector $\mathbf{c}^\alpha \triangleq [c_0^\alpha \dots c_{N_T-1}^\alpha]^T$, which is obtained from

$$\mathbf{c}^\alpha = \mathbf{R}\mathbf{b}^\alpha, \quad (3.3)$$

where \mathbf{R} is a real orthogonal matrix. Note that because of the orthogonality of \mathbf{R} , the transformation according to Eq. (3.3) preserves the Euclidean distance. A canonical representation of \mathbf{R} is given by [35, 25]

$$\mathbf{R} = \prod_{\nu=0}^{N_T-1} \prod_{\mu=\nu+1}^{N_T-1} \mathbf{T}_{\nu\mu}, \quad (3.4)$$

where the elementary rotation matrix $\mathbf{T}_{\nu\mu}$ differs from the identity matrix only in four elements: the main diagonal elements in row ν and row μ are $\cos(\phi_{\nu\mu})$, respectively, while the elements in row ν , column μ and row μ , column ν are $-\sin(\phi_{\nu\mu})$ and $\sin(\phi_{\nu\mu})$, respectively. Hence, matrix \mathbf{R} is uniquely specified by $N_T(N_T-1)/2$ phases $\phi_{\nu\mu}$ reflecting the $N_T(N_T-1)/2$ degrees of freedom of an $N_T \times N_T$ orthogonal matrix [35]. In [25] these $N_T(N_T - 1)/2$ phases are optimized for maximization of the minimum product distance.

Here, an approach more general than the one in [25] is considered. First, in general, we adopt complex signal constellations (MQAM) and, second, we allow complex unitary matrices \mathbf{R} . In particular, motivated by Eq. (3.4), we propose the construction

$$\mathbf{R} = \prod_{\nu=0}^{N_T-1} \prod_{\mu=\nu+1}^{N_T-1} \mathbf{T}_{\nu\mu}^0 \mathbf{T}_{\nu\mu}^1, \quad (3.5)$$

where $\mathbf{T}_{\nu\mu}^0$ is defined in the same way as $\mathbf{T}_{\nu\mu}$ but the corresponding phase is denoted by $\phi_{\nu\mu}^0$. $\mathbf{T}_{\nu\mu}^1$ also differs from the identity matrix only in four

⁴Similar methods have also been used for construction of modulation diversity schemes, e.g. [35, 36].

elements: the main diagonal elements in row ν and row μ , are $\cos(\phi_{\nu\mu}^1)$, respectively, while the elements in row ν , column μ and row μ , column ν are $j \sin(\phi_{\nu\mu}^1)$, respectively. The adopted construction allows for $N_T(N_T - 1)$ degrees of freedom (parameter vectors $\phi^\mu \triangleq [\phi_{01}^\mu \dots \phi_{0N_T-1}^\mu \phi_{12}^\mu \dots \phi_{N_T-2 N_T-1}^\mu]$, $\mu \in \{0, 1\}$), while general unitary matrices have N_T^2 degrees of freedom [37, Appendix D]. The “missing” N_T degrees of freedom can be realized by multiplying \mathbf{R} (as given in Eq. (3.5)) by a diagonal matrix with main diagonal entries $\exp(j\phi_\nu^2)$, $0 \leq \nu \leq N_T - 1$. However, because of the block diagonal structure of Φ_{hh} , this would not change the distance $d^2(\alpha, \beta)$ according to Eq. (2.11). For an alternative parameterization of unitary matrices we refer to [38, Appendix].

The optimum vectors ϕ_{opt}^0 and ϕ_{opt}^1 (and consequently the optimum rotation matrix \mathbf{R}_{opt}) should maximize the minimum distance d_{min}^2 (Eq. (2.12)). However, as outlined in [25] for a much simpler distance measure, a closed-form optimization seems to be not possible. Therefore, we optimized ϕ^0 and ϕ^1 by searching in discretized intervals. In order to speed up the search, we used several iterations where the interval size and the discretization size were reduced successively. The obtained parameter vectors are not necessarily the optimum ones, but, in general, they ensure a high performance.

21.3.4 Results and Discussion

In this section, we present some optimized diagonal STBC’s and discuss their performance. For this, we need to specify the channel. Here, we adopt the HT channel model [34]. For transmit and receive filtering a linearized Gaussian minimum-shift keying (GMSK) impulse [39] and a square-root raised cosine filter with roll-off factor 0.3 are adopted, respectively. This choice is motivated by the GSM/EDGE (Global System for Mobile Communications/Enhanced Data Rates for GSM Evolution) system, e.g. [30]. The discrete-time CIR was truncated to $L = 7$.

For the comparison of the optimized diagonal STBC’s we use an approximation of the BER for MLSD, which is based on a “quasi” matched filter bound for the PEP (cf. comments in Section 21.2). The approximation for BER is obtained by taking the union bound over the lower bound for PEP for all possible pairs (α, β) , $\alpha \neq \beta$, $0 \leq \alpha, \beta \leq 2^{N_T R} - 1$ [28, 29].

All HS-C’s and R-MASK’s/R-MQAM’s presented in the following were obtained as described in the previous two subsections, respectively. For all codes the labeling of the elements of the code set \mathcal{C} was optimized with respect to the adopted distance measure (Eq. (2.11)).

In Fig. 21.1 a) and b), we consider STBC’s with $N_T = 2$ and $N_T = 3$, respectively. The rate is $R = 1$ bit/(channel use) and for comparison also results for BPSK with and without receive diversity are shown. Since in Fig. 21.1

the E_b/N_0 ratio is normalized by N_R , the curves for receive diversity constitute (attainable) lower bounds for the achievable BER for transmit diversity.

For $N_T = 2$ the optimum HS-C parameter vector is $\mathbf{u}_{opt} = [1 \ 1]$, i.e., the HS-C is identical to the QPSK-RC. For R-2ASK we found the parameter vectors $\phi^0 = [0^\circ]$ and $\phi^1 = [45^\circ]$ (not necessarily optimum because of the suboptimum search method described in the previous subsection). Both HS-C and R-2ASK approach BPSK with $N_R = 2$, i.e., it can be expected that for MLSD they perform equally well. For $N_T = 3$, the optimum HS-C parameter vector is $\mathbf{u}_{opt} = [1 \ 2 \ 3]$, but the vector $\mathbf{u} = [1 \ 1 \ 3]$ yields only a slightly inferior performance. The performance of the proposed R-2ASK constellation with $\phi^0 = [22^\circ \ 27^\circ \ 11^\circ]$ and $\phi^1 = [46^\circ \ 28^\circ \ 45^\circ]$ is not much worse. On the other hand, at high signal-to-noise ratios (SNR's) 8PSK-RC imposes a performance penalty of about 3.5 dB compared to the optimum HS-C.

Diagonal STBC's with rates of $R = 2$ bits/(channel use) and $N_T = 2$ are studied in Fig. 21.2. For comparison QPSK with and without receive diversity is also considered. Now, R-4QAM with $\phi^0 = [59^\circ]$ and $\phi^1 = [45^\circ]$ yields the best performance. The optimum HS-C parameter vector is $\mathbf{u}_{opt} = [1 \ 7]$. The poorer performance of the HS-C compared to R-4QAM can be explained by the fact that for HS-C's all transmitted symbols $c_\nu[k]$ have unit magnitude which imposes additional constraints (besides the diagonal STBC matrix structure) and limits performance for higher rates. As can be observed, at high SNR's 16QAM-RC performs about 3 dB worse than R-4QAM.

A comparison of Figs. 21.1 and 21.2 reveals that the gap between receive diversity and the best diagonal STBC is larger for $R = 2$ bits/(channel use) than for $R = 1$ bit/(channel use). It can be expected that this gap further increases for higher rates and a larger number of transmit antennas. In these cases, the constraints imposed by the restriction to diagonal STBC matrices may be too stringent to yield high power efficiency.

31.3.5 Robustness of the Designed Codes

In the previous subsections, we designed codes for special fading ISI channels. A natural question is how these codes will perform if the underlying channel does not have exactly the properties we assumed for code design, which is likely to be the case in practice. Thus, in Fig. 21.3 we compare HS-C's ($R = 1$ bit/(channel use)) with $\mathbf{u} = [1 \ 1 \ 3]$ and $\mathbf{u} = [1 \ 2 \ 3]$, respectively. In Fig. 21.3a), we consider a channel whose discrete-time CIR's $h_\nu[\cdot]$, $0 \leq \nu \leq 2$, have $L = 2$ taps of equal variance. The taps are mutually correlated with correlation coefficient $r \triangleq \mathcal{E}\{h_\nu[0]h_\nu^*[1]\}/\mathcal{E}\{|h_\nu[0]|^2\}$. Our code search showed that as r tends to 1, $\mathbf{u} = [1 \ 1 \ 3]$ becomes optimum, whereas otherwise $\mathbf{u} = [1 \ 2 \ 3]$ is the optimum choice. In Fig. 21.3b), fading ISI channels with L mutually uncorrelated taps of equal variance are consid-

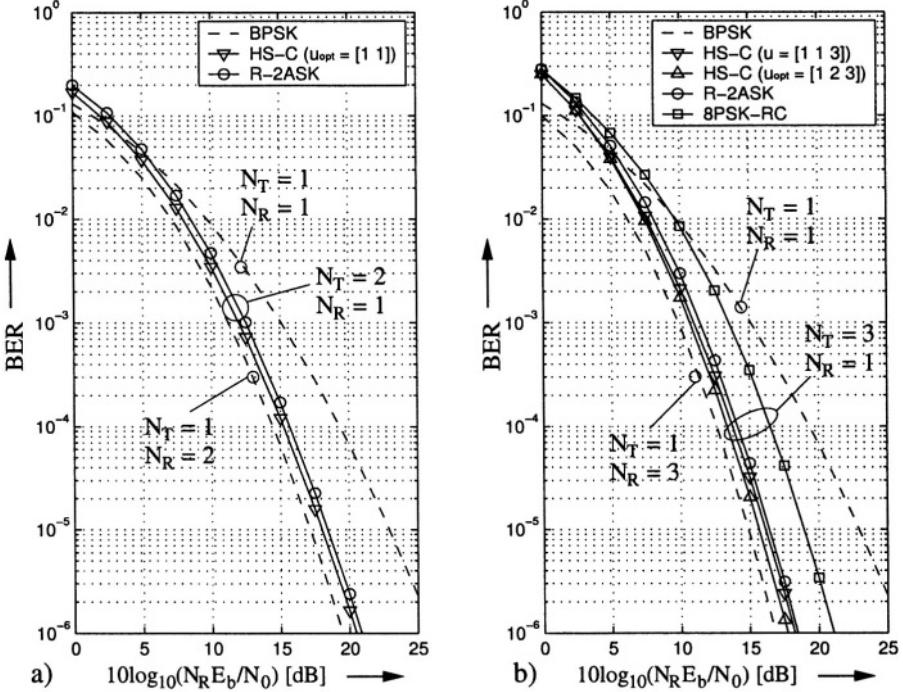


Figure 21.1. Approximate BER for MLSD vs. $10\log_{10}(N_R E_b / N_0)$ for various diagonal STBC's with a) $N_T = 2$ and b) $N_T = 3$. The rate is $R = 1$ bit/(channel use) and the HT channel is adopted. For comparison also the results for BPSK with and without receive diversity are shown.

ered. For $L = 1$ and $L > 1$ the parameter vectors $\mathbf{u} = [1 \ 1 \ 3]$ and $\mathbf{u} = [1 \ 2 \ 3]$ are optimum, respectively. Both Figs. 21.3a) and b) show that although the optimum STBC depends on the underlying channel, the designed codes are quite robust against variations of the channel characteristics. This claim is also supported by the fact that although the HS-C's in [27] were optimized for the flat fading channel (and differential encoding), for $N_T = 2$ and rates of $R = 1$ bit/(channel use) and $R = 2$ bits/(channel use) the same codes were found as in this chapter for the HT profile (and coherent transmission). Thus, in practice, we might design the STBC for a typical power delay profile and high performance can be expected even if the actual channel profile deviates to some extent from the assumed one.

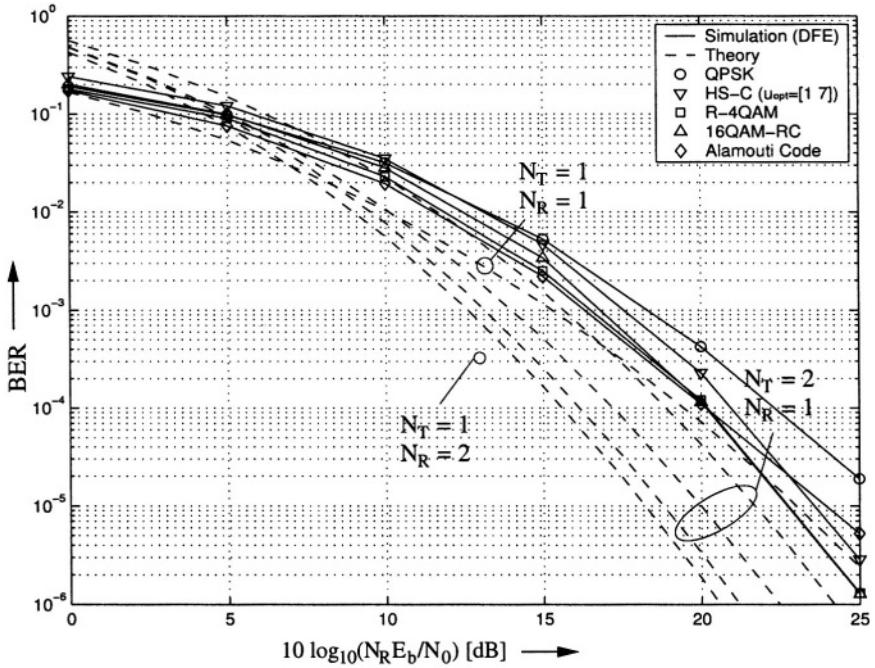


Figure 21.2. Approximate BER for MLSD vs. $10 \log_{10}(N_R E_b / N_0)$ for various diagonal STBC's with $N_T = 2$. The rate is $R = 2$ bits/(channel use) and the HT channel is adopted. For comparison also the results for QPSK with and without receive diversity are shown.

21.4 Equalization

In this section, suboptimum equalization strategies for diagonal STBC's are proposed. However, before we proceed, an equivalent $N_T \times N_T$ MIMO channel model is introduced.

21.4.1 MIMO Model

The received signal $r[k]$ according to Eq. (2.2) is cyclostationary with period $N = N_T$. Therefore, a corresponding equalizer would have to be time-variant. In order to circumvent this, we may consider the vector signal $\mathbf{r}[k] \stackrel{\Delta}{=} [r[N_T k] \ r[N_T k + 1] \ \dots \ r[N_T k + N_T - 1]]^T$ ⁵, which is stationary [40]. Taking the special structure of diagonal STBC's into account, from Eq. (2.2) we

⁵To be fully consistent with Section 21.2, $\mathbf{r}[k]$ would have to be replaced by $\mathbf{r}[k_{N_T}]$. However, in the following we simplify our notation and use k instead of k_{N_T} .

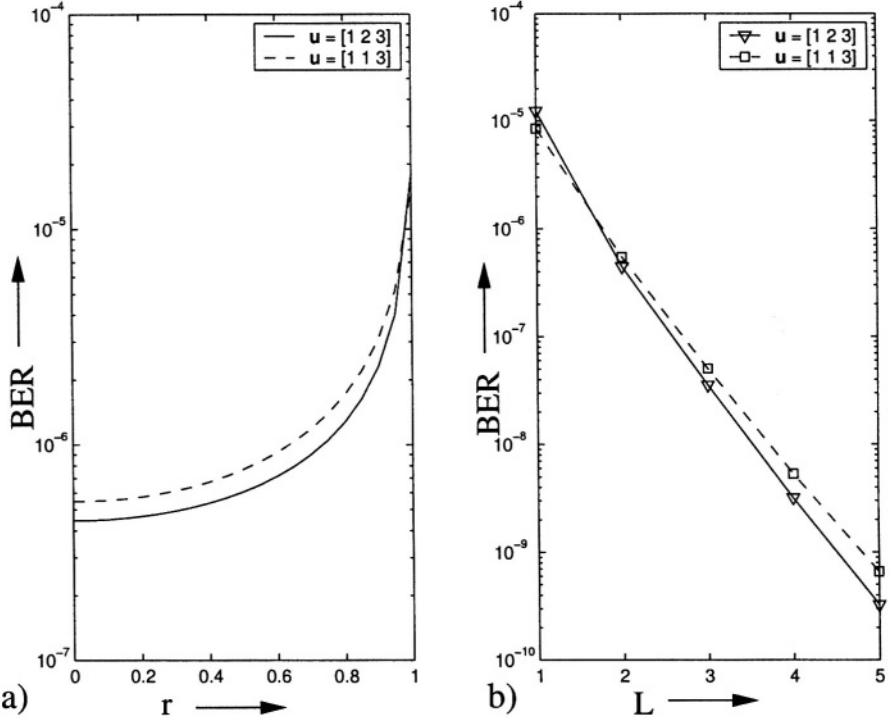


Figure 21.3. Approximate BER for MLSD vs. a) correlation coefficient r and b) L for two different HS-C's with $N_T = 3$. $R = 1$ bit/(channel use) and $10 \log_{10}(E_b/N_0) = 20$ dB is valid.

obtain

$$\mathbf{r}[k] = \sum_{l=0}^{\tilde{L}-1} \mathbf{H}[l] \mathbf{c}[k-l] + \mathbf{n}[k] \quad (4.1)$$

with the MIMO channel length $\tilde{L} \triangleq \lfloor (L-2)/N_T \rfloor + 2$ and the definitions

$$\mathbf{H}[l] \triangleq \begin{pmatrix} h_0[N_T l] & \cdots & h_{N_T-1}[N_T l - N_T + 1] \\ h_0[N_T l + 1] & \ddots & \vdots \\ \vdots & \ddots & \vdots \\ h_0[N_T l + N_T - 1] & \cdots & h_{N_T-1}[N_T l] \end{pmatrix} \quad (4.2)$$

$$\mathbf{c}[k] \triangleq [c_0[N_T k] \ c_1[N_T k + 1] \ \dots \ c_{N_T-1}[N_T k + N_T - 1]]^T \quad (4.3)$$

$$\mathbf{n}[k] \triangleq [n[N_T k] \ n[N_T k + 1] \ \dots \ n[N_T k + N_T - 1]]^T. \quad (4.4)$$

Observe that in $\mathbf{H}[\cdot]$, $h_\nu[l] = 0$ for $l < 0$ and $l \geq L$ is valid. $\mathbf{r}[\cdot]$ may be interpreted as the output of an $N_T \times N_T$ MIMO system with (matrix) impulse response $\mathbf{H}[\cdot]$. Time-invariant linear equalization (LE) and DFE of such a system is possible with mild conditions on $\mathbf{H}[\cdot]$, e.g. [40–42]. On the other hand, a general non-diagonal STBC corresponds to a MIMO system with $N_T N$ inputs and N outputs, and consequently LE and DFE lead to very poor results for $N_T > 1$.

21.4.2 Decision-Feedback Equalization

DFE appears to be a favorable equalization scheme for the above MIMO system. In a MIMO DFE the vector signal $\mathbf{y}[k]$ at the input of the decision device is given by (cf. Fig. 21.4)

$$\mathbf{y}[k] = \sum_{l=0}^{L_F-1} \mathbf{W}[l] \mathbf{r}[k-l] - \sum_{l=1}^{L_B} \mathbf{B}[l] \hat{\mathbf{c}}[k-l-k_0], \quad (4.5)$$

where $\mathbf{W}[l]$, $0 \leq l \leq L_F - 1$, and $\mathbf{B}[l]$, $1 \leq l \leq L_B$, are the $N_T \times N_T$ matrix filter coefficients of the feedforward (FF) and the feedback (FB) filter, respectively. L_F , L_B , and k_0 are the FF filter length, the FB filter length, and a suitably chosen decision delay, respectively. $\hat{\mathbf{c}}[\cdot]$ denotes the estimate for $\mathbf{c}[\cdot]$. Various methods for optimization of the FF and FB filters can be found in the literature, e.g. [40–42]. In this chapter, we adopt the finite impulse response (FIR) minimum mean-squared error (MMSE) MIMO DFE proposed in [41, Scenario 1]. This approach has the advantage that fast algorithms for closed-form filter calculation are available [41] (for this, the channel noise variance σ_n^2 and the CIR coefficients $h_\nu[l]$, $0 \leq l \leq L - 1$, $0 \leq \nu \leq N_T - 1$, have to be known) and alternatively, the filters may be calculated recursively using adaptive algorithms, e.g. [43].

If the FF and FB filters are adjusted properly, $\mathbf{y}[k]$ can be modeled as

$$\mathbf{y}[k] = \mathbf{c}[k - k_0] + \mathbf{e}[k], \quad (4.6)$$

where $\mathbf{e}[\cdot]$ denotes the error signal vector. Although it can be expected that $\mathbf{e}[k]$ is a temporally white process provided that the FF and FB filters are sufficiently long [44], in general, for a given k the elements of $\mathbf{e}[k]$ are mutually correlated. The error autocorrelation matrix

$$\Phi_{ee} \stackrel{\Delta}{=} \mathcal{E}\{\mathbf{e}[k]\mathbf{e}^H[k]\} \quad (4.7)$$

can be calculated in closed-form from [41, Eq. (16)] or estimated using standard techniques, e.g. [45].

Although $\mathbf{e}[k]$ is not exactly Gaussian distributed, in general, for long FF and FB filters its probability density function (pdf) may be well approximated

as a multivariate Gaussian distribution. Consequently, the vector $\mathbf{c}[k - k_0]$ which contains the diagonal elements of the transmitted STBC matrix may be obtained from

$$\hat{\mathbf{c}}[k - k_0] = \underset{\mathbf{c}[k-k_0]}{\operatorname{argmin}} \left\{ (\mathbf{y}[k] - \mathbf{c}[k - k_0])^H \Phi_{ee}^{-1} (\mathbf{y}[k] - \mathbf{c}[k - k_0]) \right\}. \quad (4.8)$$

In order to reduce the computational complexity, the *Cholesky factorization*

$$\Phi_{ee}^{-1} \triangleq \mathbf{L}^H \Phi_{dd}^{-1} \mathbf{L} \quad (4.9)$$

may be used. Here, \mathbf{L} is a lower triangular matrix, whose main diagonal elements are all ones, and Φ_{dd} is a diagonal matrix with main diagonal entries d_ν^2 , $0 \leq \nu \leq N_T - 1$. If we precompute

$$\begin{aligned} \bar{\mathbf{y}}[k] &\triangleq [\bar{y}[N_T k] \dots \bar{y}[N_T k + N_T - 1]]^T \\ &\triangleq \mathbf{L} \mathbf{y}[k] \end{aligned} \quad (4.10)$$

and all possible vectors

$$\begin{aligned} \bar{\mathbf{c}}[k] &\triangleq [\bar{c}_0[N_T k] \dots \bar{c}_{N_T-1}[N_T k + N_T - 1]]^T \\ &\triangleq \mathbf{L} \mathbf{c}[k], \end{aligned} \quad (4.11)$$

Eq. (4.8) simplifies to

$$\hat{\mathbf{c}}[k - k_0] = \underset{\bar{\mathbf{c}}[k-k_0]}{\operatorname{argmin}} \left\{ \sum_{\nu=0}^{N_T-1} \frac{1}{d_\nu^2} |\bar{y}[N_T k + \nu] - \bar{c}_\nu[N_T(k - k_0) + \nu]|^2 \right\}, \quad (4.12)$$

where the estimate for the original signal vector can be obtained from $\hat{\mathbf{c}}[k - k_0] = \mathbf{L}^{-1} \hat{\mathbf{c}}[k - k_0]$.

Another (suboptimum) possibility to reduce computational complexity in Eq. (4.8) is to neglect the mutual correlations of the elements of $\mathbf{e}[\cdot]$. In this case, also Eq. (4.12) can be used for determination of $\hat{\mathbf{c}}[k - k_0]$, however, now $\mathbf{L} \triangleq \mathbf{I}_{N_T}$ is valid and d_ν^2 has to be replaced by the ν th main diagonal element of the error ACM Φ_{ee} . The associated total error variance relevant for the performance of the MIMO DFE is

$$\sigma_e^2 \triangleq \operatorname{tr}\{\Phi_{ee}\}, \quad (4.13)$$

whereas the relevant error variance for the first approach is

$$\sigma_d^2 \triangleq \operatorname{tr}\{\Phi_{dd}\}. \quad (4.14)$$

Since the Cholesky factorization accomplishes the (spatial) whitening of $\mathbf{e}[k]$, i.e., the components of $\mathbf{L}\mathbf{e}[k]$ are mutually uncorrelated,

$$\sigma_d^2 \leq \sigma_e^2 \quad (4.15)$$

is valid (with equality if and only if the components of $\mathbf{e}[k]$ are mutually uncorrelated, e.g. [45]).

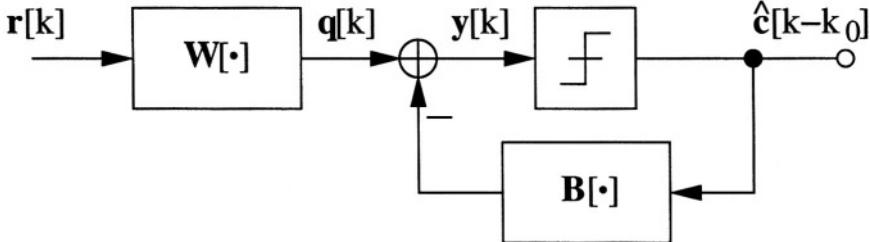


Figure 21.4. Block diagram of receiver structure for MIMO DFE.

21.4.3 Decision–Feedback Sequence Estimation (DFSE)

The performance gap between DFE and MLSD can be bridged by RSSE techniques at the expense of additional complexity [15]. Here, we focus on DFSE [16]. In general, DFSE can yield high performance only if the overall MIMO channel transfer function is minimum phase, i.e., an appropriate prefilter has to be applied to transform the original MIMO channel transfer function

$$\mathbf{H}(z) \triangleq \sum_{l=0}^{L-1} \mathbf{H}[l] z^{-l} \quad (4.16)$$

into its minimum-phase equivalent $\mathbf{H}_{\min}(z)$. Similar to the single-input single-output (SISO) case [46], also for the MIMO case the FIR–MMSE–DFE FF filter appears to be a favorable choice for the prefilter since it accomplishes an approximate minimum-phase transformation and can be computed easily.

For DFSE the absolute metric

$$\Lambda[k+1] = \Lambda[k] + \lambda[k] \quad (4.17)$$

can be minimized by applying the Viterbi algorithm (VA) [47, 48] on a trellis with $Z = 2^{N_T R K}$ states. The design parameter K , $0 \leq K \leq L_B$, allows to adjust the desired trade-off between complexity and power efficiency. The DFSE branch metric $\lambda[k]$ is given by

$$\lambda[k] = \left(\mathbf{q}[k] - \sum_{l=0}^K \mathbf{B}[l] \hat{\mathbf{c}}[k - k_0 - l] - \sum_{l=K+1}^{L_B} \mathbf{B}[l] \hat{\mathbf{c}}[k - k_0 - l] \right)^H \boldsymbol{\Phi}_{ee}^{-1}$$

$$\left(\mathbf{q}[k] - \sum_{l=0}^K \mathbf{B}[l] \tilde{\mathbf{c}}[k - k_0 - l] - \sum_{l=K+1}^{L_B} \mathbf{B}[l] \hat{\mathbf{c}}[k - k_0 - l] \right) \quad (4.18)$$

where

$$\mathbf{q}[k] = \sum_{l=0}^{L_F-1} \mathbf{W}[l] \mathbf{r}[k - l] \quad (4.19)$$

is the output of the DFE FF filter and the definition $\mathbf{B}[0] \triangleq \mathbf{I}_{N_T}$ is used, $\tilde{\mathbf{c}}[\cdot]$ and $\hat{\mathbf{c}}[\cdot]$ denote equalizer trial symbols (defined by the states and the transitions of the trellis) and state-dependent feedback symbols, respectively. Note that for the limiting cases $K = 0$ and $K = L_B$ DFSE is identical to DFE and MLSD, respectively. An alternative representation of the branch metric $\lambda[k]$, analogous to Eq. (4.12), is straightforward.

21.4.4 Suitability of STBC for Suboptimum Equalization

The optimality criterion adopted for code design in Section 21.3 guarantees high performance for MLSD. Since the power efficiency of suboptimum equalizers is bounded by that of optimum MLSD and the applicability of suboptimum equalizers is facilitated by the diagonal structure of the code matrices, this is a reasonable approach. However, suppose there are two STBC's which have a similar performance for MLSD, which one should we adopt if suboptimum equalization is to be used?

Let us first consider DFE. Here, the receiver performance is not only governed by the distance properties of the STBC but also by the total error variance σ_d^2 . On the other hand, as can be observed from [41, Eqs. (14), (16)], the total error variance depends on the ACM $\Phi_{cc} \triangleq \mathcal{E}\{\mathbf{c}[k]\mathbf{c}^H[k]\}$ of the MIMO input vector $\mathbf{c}[k]$, i.e., on the second-order statistics of the STBC signal set. In order to illustrate this dependence, we consider the special cases

$$\Phi_{cc} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \quad (4.20)$$

and

$$\Phi_{cc} = \begin{pmatrix} 1 & \rho & 0 \\ \rho & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad (4.21)$$

for $N_T = 2$ and $N_T = 3$, respectively, where ρ , $0 \leq \rho \leq 1$, is the correlation coefficient. In Figs. 21.5a) and b) we depict the corresponding average total error variance $\overline{\sigma_d^2}$ vs. ρ for $10 \log_{10}(E_b/N_0) = 10$ dB and $10 \log_{10}(E_b/N_0) = 20$ dB, respectively. Thereby, $L_F = 10$, $L_B = 3$ and $L_F = 7$, $L_B = 2$ has been used for $N_T = 2$ and $N_T = 3$, respectively. In addition, $R =$

1 bit/(channel use) is valid and $\overline{\sigma_d^2}$ is obtained by averaging over 1000 realizations of an HT channel ($L = 7$). For comparison, also the average total error variance $\overline{\sigma_e^2}$ is shown, which would be relevant if spatial noise whitening was not performed. From Fig. 21.5 we observe that the total error variance can be considerably reduced by spatial noise whitening for both $N_T = 2$ and $N_T = 3$, and for all values of ρ . Another very interesting observation is that both $\overline{\sigma_d^2}$ and $\overline{\sigma_e^2}$ decrease with increasing correlation coefficient ρ . This suggests that correlations among the signals transmitted from different antennas are beneficial for DFE. Indeed, we will see in Section 21.5 that DFE can approach MLSD closely for STBC's with strong (spatial) correlations, whereas a loss in performance of about 3–4 dB results if there are no correlations. As an example, we may consider the HS-C and R-2ASK codes discussed in Section 21.3 for $N_T = 2$ and $R = 1$ bit/(channel use), which yield a similar performance for MLSD, cf. Fig. 21.1a). It can be easily checked that the ACM's for the HS-C and the R-2ASK are given by Eq. (4.20) with $\rho = 1$ and $\rho = 0$, respectively. As expected, for DFE the HS-C outperforms the R-2ASK by about 3 dB. For DFSE this gap reduces with increasing K . As a second example, we consider the two HS-C's proposed in Section 21.3 for $N_T = 3$ and $R = 1$ bit/(channel use). The ACM's for $\mathbf{u} = [1 \ 1 \ 3]$ and $\mathbf{u} = [1 \ 2 \ 3]$ are given by Eqs. (4.21) with $\rho = 1$ and $\rho = 0$, respectively. Although Fig. 21.1b) suggests that $\mathbf{u} = [1 \ 2 \ 3]$ yields a superior performance for MLSD, our simulations showed that $\mathbf{u} = [1 \ 1 \ 3]$ performs better for DFE.

The above findings can be explained as follows. If the symbols emitted by the transmit antennas are mutually uncorrelated, the DFE, which relies on second-order statistics only, treats them as originating from independent sources and consequently has to perform interference suppression to separate them. On the other hand, if the transmitted symbols are strongly mutually correlated, the DFE essentially can combine the contributions originating from different transmit antennas and thus, can yield a lower total error variance.

21.4.5 Complexity Issues

In this section, we briefly discuss the complexity of the proposed receivers and compare it with that of SISO equalization and with that of the equalizers reported in [12]. A comparison with other receivers for STBC's, e.g. [6, 9], seems to be difficult since these schemes require more than one receive antenna.

For the MIMO DFE scheme presented here, there are $N_T^2(L_F + L_B)$ scalar FF and FB filter taps, and the decision rules according to Eqs. (4.8), (4.12) demand $2^{N_T R}/(N_T R)$ metric calculations per decided bit. For the SISO case only $L_F + L_B$ filter taps and $2^R/R$ metric calculations per bit decision are necessary. However, if we take into account that for long SISO channels (i.e.,

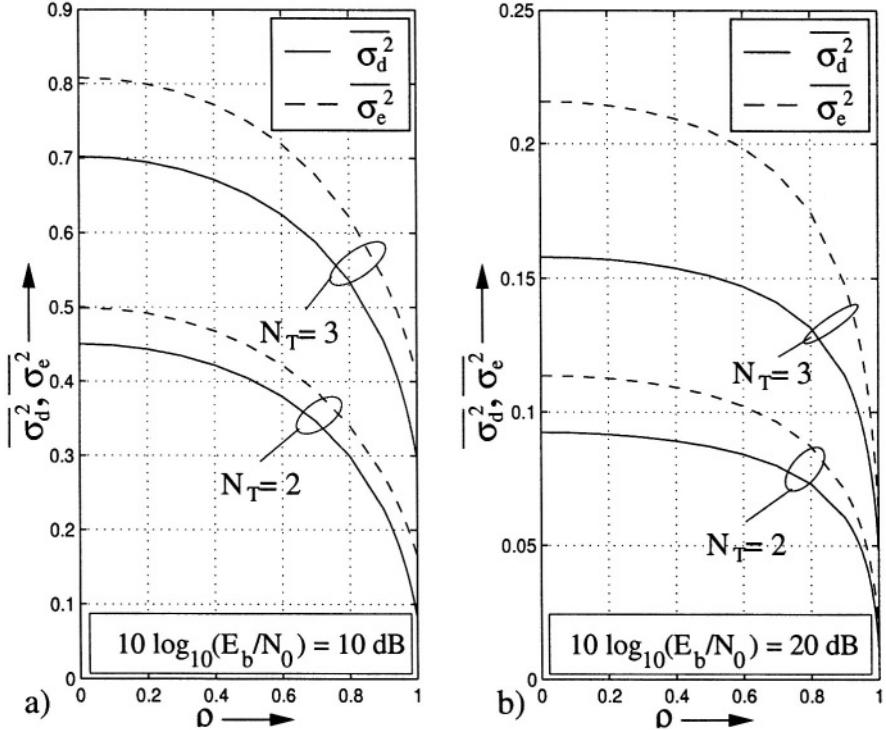


Figure 21.5. Average total error variances $\overline{\sigma}_d^2$ and $\overline{\sigma}_e^2$ vs. ρ for a) $10 \log_{10}(E_b/N_0) = 10$ dB and b) $10 \log_{10}(E_b/N_0) = 20$ dB. The HT channel model is adopted and $R = 1$ bit/(channel use) is valid.

$L \gg N_T$) the effective channel length of the corresponding MIMO channel is $\tilde{L} \approx L/N_T$, and consequently, the MIMO DFE FF and FB filters can be made by a factor of N_T shorter than the SISO DFE filters, the number of scalar filter taps for the ST coded system is approximately by a factor of N_T larger than for single antenna transmission.

For the scheme in [12], cf. also [22, 23], which is tailored for Alamouti's code (i.e., $N_T = 2$), not only the received signal vector $\mathbf{r}[k]$ but also $\mathbf{r}^*[k]$ is processed. Thus, $(2N_T)^2(L_F + L_B)$ scalar FF and FB filter taps have to be calculated. If symmetry relations are exploited, half of the calculated taps can be discarded and $2N_T^2(L_F + L_B)$ taps are necessary for filtering. Due to the orthogonal structure of Alamouti's code, only $2^R/R$ metric calculations per bit decision have to be carried out. For low rates and a small number of transmit antennas, the approach presented in this chapter requires a lower computational complexity than the scheme presented in [12] because of the

lower number of filter taps. For higher rates the number of metric calculations becomes more important and the scheme in [12] has a complexity advantage. However, it should be noted that diagonal STBC's are mainly beneficial for rates $R \leq 2$ bits/(channel use) anyway.

Finally, it is worth mentioning that for both the DFSE scheme proposed here for diagonal STBC's and that proposed in [12] for Alamouti's code $Z = 2^{N_T R K}$ states are necessary if a (MIMO) channel memory of K symbols is taken into account, whereas for SISO DFSE only $Z = 2^{R K}$ states are required.

21.4.6 Extensions

There are several useful extensions to the proposed suboptimum equalizers. As a first example, instead of simple DFSE more sophisticated RSSE techniques relying on set partitioning [15] might be used. It can be expected that similar to the SISO case, e.g. [30], this would allow to reduce complexity without compromising performance. Furthermore, as an alternative approach to RSSE the MIMO model presented here also enables the application of impulse response truncation techniques [49].

In this work, we restrict our attention to coherent equalization. Nevertheless, since HS-C's can be differentially encoded [27], an interesting field for future research is the design of noncoherent equalizers for ST block coded transmission. Similar to differential PSK transmission in the single antenna case, e.g. [50, 51], it is expected that noncoherent equalizers would enable a favorable trade-off between power efficiency and robustness against phase noise and frequency offsets.

21.5 Simulation Results

In this section, we present some simulation results for the diagonal STBC's proposed in Section 21.3 and the suboptimum equalization schemes discussed in Section 21.4. We adopt the transmit and receiver input filters as described in Section 21.3, and for Figs. 21.6–21.8 the HT channel model is used, whereas for Fig. 21.9 the typical urban (TU) channel model [34] is valid. For the HT channel ($L = 7$) we adopted $L_F = 20$ ($L_B = 6$), $L_F = 10$ ($L_B = 3$), and $L_F = 7$ ($L_B = 2$), for $N_T = 1$, $N_T = 2$, and $N_T = 3$, respectively, whereas for the TU channel ($L = 5$) we adopted $L_F = 20$ ($L_B = 4$), $L_F = 10$ ($L_B = 2$), and $L_F = 7$ ($L_B = 1$), for $N_T = 1$, $N_T = 2$, and $N_T = 3$, respectively. For all simulations perfect knowledge of the CIR's and the noise variance at the receiver is assumed.

In Fig. 21.6 BER vs. $10 \log_{10}(E_b/N_0)$ is depicted for $R = 1$ bit/(channel use) and BPSK ($N_T = 1$), HS-C's ($N_T = 2$ and $N_T = 3$), and Alamouti's code ($N_T = 2$). Besides the simulation results for DFE and DFSE also the approximation for the BER for MLSD (dashed lines) is shown, cf. Section 21.3,

[28, 29]. For the reasons discussed in Section 21.4 for $N_T = 2$ we consider the HS-C rather than R-2ASK, and for $N_T = 3$ the HS-C parameter vector $\mathbf{u} = [1 \ 1 \ 3]$ is adopted instead of $\mathbf{u}_{opt} = [1 \ 2 \ 3]$, which is optimum for MLSD but has an inferior performance for DFE. Indeed, it can be observed that for $N_T = 2$ the HS-C with DFE suffers only a small performance penalty of 0.5 dB compared to the approximation for MLSD. For $N_T = 3$ the gap for DFE is about 1 dB but can be narrowed by DFSE with $Z = 8$ states ($K = 1$). On the other hand, Alamouti's code suffers from a large loss in performance if widely linear (WL) DFE [12, 22] is adopted instead of MLSD. The BER curve for WL-DFE is essentially parallel to the curves for BPSK, i.e., it appears that Alamouti's code cannot take advantage of the additional diversity offered by two transmit antennas if suboptimum equalizers are used. This might be attributed to the additional interference which is caused if two antennas transmit simultaneously and which has to be suppressed by the WL-DFE.

Fig. 21.7 shows simulation results for DFE for QPSK, Alamouti's code, and the diagonal STBC's already considered in Fig. 21.2. The rate is $R = 2$ bits/(channel use) and for comparison also the corresponding approximate BER's for MLSD are depicted. Although at high SNR's the approximate BER for MLSD for 16QAM-RC is about 3 dB worse than that for R-4QAM, both STBC's have a similar performance if DFE is applied. This behavior can again be explained by the findings in Section 21.4 since the symbols transmitted over the two antennas are fully mutually correlated and uncorrelated for 16QAM-RC and R-4QAM, respectively. At low SNR's Alamouti's code with WL-DFE performs as good R-4QAM with DFE, whereas the diagonal STBC's perform better at high SNR's, where the curve for Alamouti's code is again essentially parallel to that for $N_T = 1$ (QPSK). At $\text{BER} = 10^{-5}$, R-4QAM outperforms Alamouti's code and QPSK by 1.3 dB and 3.8 dB, respectively. As can be observed from Fig. 21.8, the gap between DFE and MLSD can be closed by application of DFSE for both HS-C and R-4QAM at the expense of an increase in computational complexity. It is also interesting to note that the performance gap between different diagonal STBC's for DFE and DFSE can be accurately predicted by the performance gap of the respective approximate BER curves for MLSD as long as the STBC's have the same correlation properties.

So far, we considered only the HT channel model. In order to illustrate the possible performance gains for a different power delay profile, we adopt the TU channel model for Fig. 21.9. For TU and $R = 1$ bit/(channel use) the same HS-C's as for HT are optimum. Again, we consider the parameter vector $\mathbf{u} = [1 \ 1 \ 3]$ instead of $\mathbf{u}_{opt} = [1 \ 2 \ 3]$ since this leads to a better performance for DFE. Fig. 21.9 shows that for the TU channel both the HS-C for $N_T = 2$ and that for $N_T = 3$ perform close to the approximate BER for MLSD. The small gap between DFE and MLSD for $N_T = 3$ can be attributed to short feedback filter length ($L_B = 1$ is valid). Alamouti's code with WL-DFE is

clearly outperformed by the diagonal STBC's and, at high SNR's, it performs even worse than BPSK.

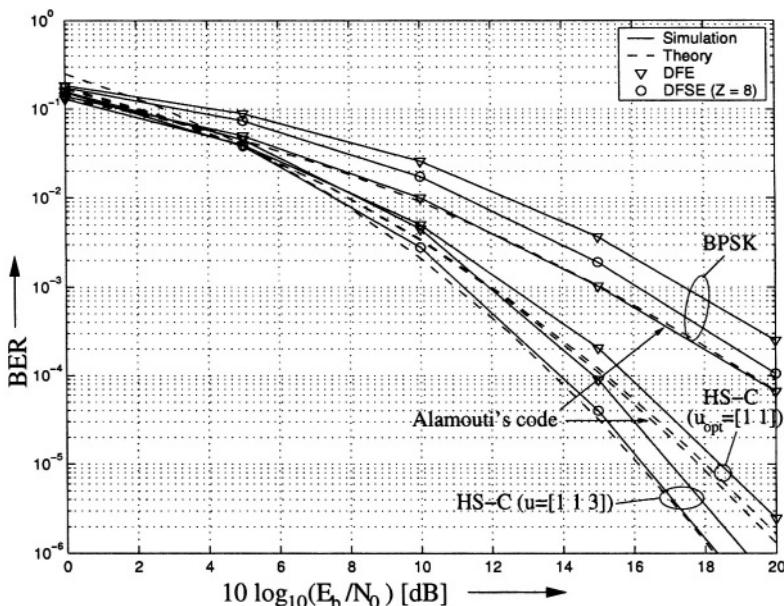


Figure 21.6. BER vs. $10 \log_{10}(E_b/N_0)$ for BPSK ($N_T = 1$), HS-C's ($N_T = 2$ and $N_T = 3$), and Alamouti's code ($N_T = 2$). The rate is $R = 1$ bit/(channel use) and the HT channel is adopted.

21.6 Conclusions

In this chapter, diagonal STBC's for transmission over fading ISI channels have been optimized based on a bound for the PEP for MLSD and their performance has been studied. An equivalent MIMO channel model enabled the successful application of suboptimum equalization techniques although (only a single receive antenna was employed. It has been shown that for DFE also the mutual correlation of the symbols transmitted over different antennas plays an important role but becomes less important for DFSE as the number of states is increased. Simulations have shown that diagonal STBC's with DFE clearly outperform Alamouti's code with WL-DFE [12, 22] for rates $R \leq 2$ bit/(channel use). For higher rates the power efficiency of diagonal STBC's decreases due to the constraints imposed by the restriction to diagonal code matrices. Here, the design of more power efficient STBC's allowing for suboptimum equalization is a promising field for future research. Thereby, a similar approach as in this chapter may be adopted, i.e., the general STBC

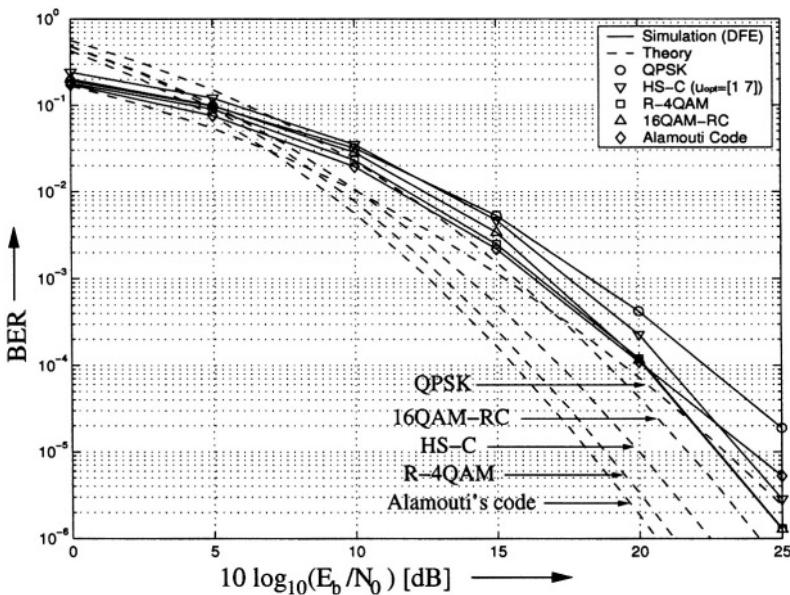


Figure 21.7. BER vs. $10 \log_{10}(E_b/N_0)$ for QPSK ($N_T = 1$) and various STBC's ($N_T = 2$). The rate is $R = 2$ bits/(channel use) and the HT channel is adopted.

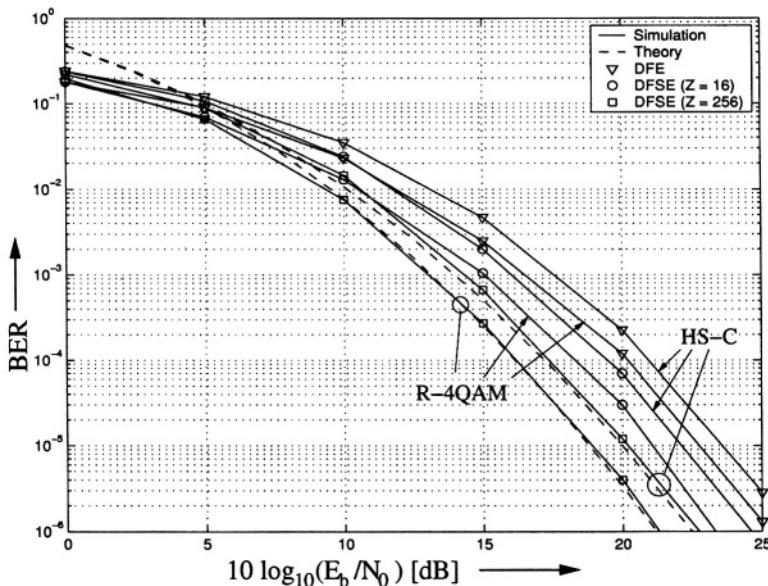


Figure 21.8. BER vs. $10 \log_{10}(E_b/N_0)$ for HS-C and R-4QAM with DFE and DFSE. The rate is $R = 2$ bits/(channel use) and the HT channel is adopted.

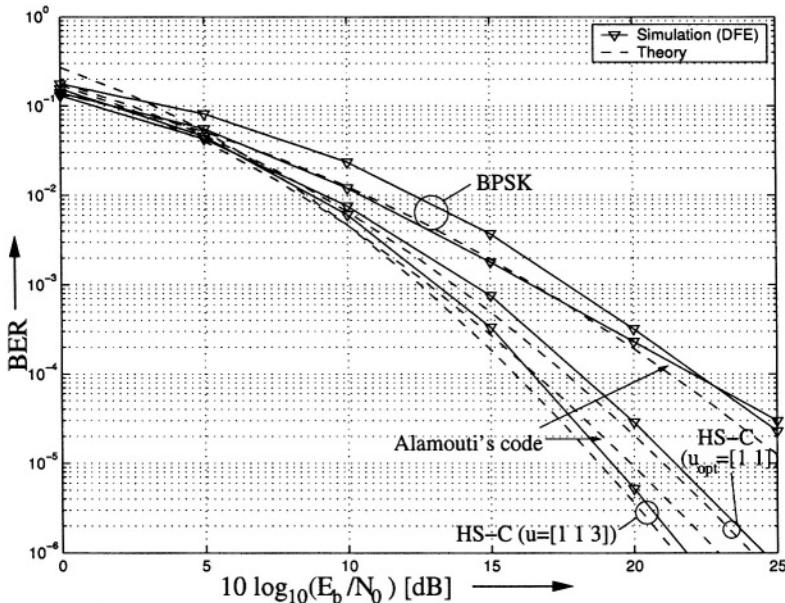


Figure 21.9. BER vs. $10 \log_{10}(E_b/N_0)$ for BPSK ($N_T = 1$), HS-C's ($N_T = 2$ and $N_T = 3$), and Alamouti's code ($N_T = 2$). The rate is $R = 1$ bit/(channel use) and the TU channel is adopted.

design criterion can be used while constraints on the structure of the STBC matrix ensure the applicability of suboptimum (WL) equalization.

References

- [1] A. Wittneben. A New Bandwidth Efficient Transmit Antenna Modulation Diversity Scheme for Linear Digital Modulation. In *Proceedings of the International Conference on Communications*, pages 1630–1634, Geneva, May 1993.
- [2] S.M. Alamouti. A Simple Transmitter Diversity Scheme for Wireless Communications. *IEEE Journal on Selected Areas in Communications*, SAC-16:1451–1458, October 1998.
- [3] V. Tarokh, N. Seshadri, and A.R. Calderbank. Space-Time Codes for High Data Rate Wireless Communication: Performance Criterion and Code Construction. *IEEE Transactions on Information Theory*, IT-44:744–765, March 1998.
- [4] J.-C. Guey, M.P. Fitz, M.R. Bell, and W.-Y. Kuo. Signal Design for Transmitter Diversity Wireless Communication Systems Over Rayleigh Fading Channels. *IEEE Transactions on Communications*, COM-47:527–537, April 1999.

- [5] D. Agrawal, V. Tarokh, A. Naguib, and N. Seshadri. Space–Time Coded OFDM for High Data–Rate Wireless Communication Over Wideband Channels. In *Proceedings of the IEEE Vehicular Technology Conference*, pages 2232–2236, Ottawa, May 1998.
- [6] W.-J. Choi and J.M. Cioffi. Space–Time Block Codes over Frequency Selective Rayleigh Fading Channels. In *IEEE Vehicular Technology Conference*, pages 2541–2545, November 1999.
- [7] Y. Li, J.C. Chung, and N.R. Sollenberger. Transmitter Diversity for OFDM Systems and its Impact on High–Rate Data Wireless Networks. *IEEE Journal on Selected Areas in Communications*, SAC-17:1233–1243, July 1999.
- [8] E. Lindskog and A. Paulraj. A transmit diversity scheme for channels with intersymbol interference. In *Proceedings of the IEEE International Conference on Communications*, New Orleans, June 2000.
- [9] N. Al-Dahir, A.F. Naguib, and A.R. Calderbank. Finite–Length MIMO Decision–Feedback Equalization for Space–Time Block–Coded Signals Over Multipath–Fading Channels. *IEEE Transactions on Vehicular Technology*, VT-50:1176–1182, July 2001.
- [10] N. Al-Dahir. Single–Carrier Frequency–Domain Equalization for Space–Time Block–Coded Transmission Over Frequency–Selective Fading Channels. *IEEE Communications Letters*, 5:304–306, July 2001.
- [11] Z. Liu, G.B. Giannakis, S. Barbarossa, and A. Scaglione. Transmit–Antennae Space–Time Block Coding for Generalized OFDM in the Presence of Unknown Multipath. *IEEE Transactions on Communications*, COM-49:1352–1364, July 2001.
- [12] W.H. Gerstacker, F. Obernosterer, R. Schober, A. Lehmann, A. Lampe, and P. Gunreben. Receiver Concepts for Space–Time Block–Coded Transmission over Frequency–Selective Fading Channels. *Revised version submitted to IEEE Transactions on Communications*, September 2003.
- [13] L. Li, H. Li, and Y.-D. Yao. Transmit Diversity and Equalization for Frequency Selective Fading Channels. In *IEEE Vehicular Technology Conference*, pp. 1673–1677, Rhodes, Greece, May 2001.
- [14] S. Zhou and G.B. Giannakis. Single–Carrier Space–Time Block–Coded Transmission Over Frequency–Selective Fading Channels. *IEEE Transactions on Information Theory*, IT-49:164–179, January 2003.
- [15] M.V. Eyuboglu and S.U. Qureshi. Reduced–State Sequence Estimation with Set Partitioning and Decision Feedback. *IEEE Transactions on Communications*, COM-36:13–20, January 1988.

- [16] A. Duel-Hallen and A. C. Heegard. Delayed Decision–Feedback Sequence Estimation. *IEEE Transactions on Communications*, COM-37:428–436, 1989.
- [17] M.J. Heikkilae, K. Majonen, and J. Lilleberg. Decoding and Performance of Space–Time Trellis Codes in Fading Channels with Intersymbol Interference. In *Proceedings of the IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, pages 1077–1082, London, September 2000.
- [18] A.F. Naguib. Equalization of Transmit Diversity Space–Time Coded Signals. In *IEEE Global Telecommunications Conference*, San Francisco, December 2000.
- [19] W. Younis and N. Al-Dahir. Joint Prefiltering and MLSE Equalization of Space–Time–Coded Transmission Over Frequency–Selective Channels. *IEEE Transactions on Vehicular Technology*, VT-51:144–154, January 2002.
- [20] V. Tarokh, H. Jafarkhani, and A.R. Calderbank. Space–Time Block Codes from Orthogonal Designs. *IEEE Transactions on Information Theory*, IT-45:1456–1467, July 1999.
- [21] N. Al-Dahir. Overview and Comparison of Equalization Schemes for Space–Time–Coded Signals With Application to EDGE. *IEEE Transactions on Signal Processing*, SP-50:2477–2488, October 2002.
- [22] W.H. Gerstacker, F. Obernosterer, R. Schober, A. Lehmann, A. Lampe, and P. Gunreben. Widely Linear Equalization for Space–Time Block–Coded Transmission Over Fading ISI Channels. In *Proceedings of the IEEE Vehicular Technology Conference*, pages 238–242, Vancouver, Canada, September 2002.
- [23] W.H. Gerstacker, F. Obernosterer, R. Schober, A. Lehmann, A. Lampe, and P. Gunreben. Symbol–by–Symbol and Trellis–Based Equalization with Widely Linear Processing for Space–Time Block–Coded Transmission over Frequency–Selective Channels. In *Proceedings of the IEEE Global Telecommunications Conference*, Taipei, Taiwan, December 2002.
- [24] B. Picinbono and P. Chevalier. Widely linear estimation with complex data. *IEEE Transactions on Signal Processing*, 43:2030–2033, August 1995.
- [25] V.M. DaSilva and E.S. Sousa. Fading–Resistant Modulation Using Several Transmitter Antennas. *IEEE Transactions on Communications*, COM-45:1236–1244, October 1997.
- [26] B.L. Hughes. Differential Space–Time Modulation. *IEEE Transactions on Information Theory*, IT-46:2567–2578, November 2000.

- [27] B.M. Hochwald and W. Sweldens. Differential Unitary Space–Time Modulation. *IEEE Transactions on Communications*, COM-48:2041–2052, December 2000.
- [28] R. Schober, W. H. Gerstacker, and L.H.-J. Lampe. Performance Analysis and Design of STBC’s for Frequency–Selective Fading Channels. *To appear in IEEE Transactions on Wireless Communications*, 2003.
- [29] R. Schober, W. H. Gerstacker, and L.H.-J. Lampe. Performance Analysis and Design of STBC’s for Fading ISI Channels. *Proceedings of the IEEE International Conference on Communications*, pp. 1451–1455, New York, May 2002.
- [30] W. H. Gerstacker and R. Schober. Equalization Concepts for EDGE. *IEEE Transactions on Wireless Communications*, TW-01:190–199, January 2002.
- [31] M.V. Clark, L.J. Greenstein, W.K. Kennedy, and M. Shafi. Matched Filter Performance Bounds for Diversity Combining Receivers in Digital Mobile Radio. *IEEE Transactions on Vehicular Technology*, VT-41:356–362, November 1992.
- [32] Y. Liu, M.P. Fitz, and O.Y. Takeshita. Space–time codes performance criteria and design for frequency selective fading channels. In *Proceedings of the IEEE International Conference on Communications*, Helsinki, June 2001.
- [33] N.W.K. Lo, D.D. Falconer, and A.U.H. Sheikh. Adaptive Equalization for Co-Channel Interference in a Multipath Fading Environment. *IEEE Transactions on Communications*, COM-43:787–794, February–April 1995.
- [34] *GSM Recommendation 05.05: “Propagation Conditions”, Vers. 5.3.0, Release 1996.*
- [35] D. Rainish. Diversity Transform for Fading Channels. *IEEE Transactions on Communications*, COM-44:1653–1661, December 1996.
- [36] J. Boutros and E. Viterbo. Signal Space Diversity: A Power- and Bandwidth-Efficient Diversity Technique for the Rayleigh Fading Channel. *IEEE Transactions on Information Theory*, IT-44:1453–1467, July 1998.
- [37] P.P. Vaidyanathan, T.Q. Nguyen, Z. Doganata, and T. Saramaki. Improved Technique for Design of Perfect Reconstruction FIR QMF Banks with Lossless Polyphase Matrices. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37:1042–1056, July 1989.
- [38] D. Agrawal, T.J. Richardson, and R.L. Urbanke. Multiple–Antenna Signal Constellations for Fading Channels. *IEEE Transactions on Information Theory*, IT-47:2618–2626, September 2001.
- [39] *ETSI, Tdoc SMG2 WPB 325/98, November 1998.*

- [40] A. Duel-Hallen. Equalizers for Multiple Input/Multiple Output Channels and PAM Systems with Cyclostationary Input Sequences. *IEEE Journal on Selected Areas in Communications*, SAC-10:630–639, April 1992.
- [41] N. Al-Dhahir and A.H. Sayed. The Finite-Length Multi-Input Multi-Output MMSE-DFE. *IEEE Transactions on Signal Processing*, SP-48:2921–2936, October 2000.
- [42] C. Tidestav, A. Ahlen, and M. Sternad. Realizable MIMO Decision Feedback Equalizers: Structure and Design. *IEEE Transactions on Signal Processing*, SP-49:121–133, January 2001.
- [43] A. Maleki-Tehrani, B. Hassibi, and J.M. Cioffi. Adaptive Equalization of Multiple-Input Multiple-Output (MIMO) Channels. In *IEEE International Conference on Communications*, pages 1670–1674, New Orleans, June 2000.
- [44] J. M. Cioffi, G. P. Dudevoir, M. V. Eyuboglu, and G. D. Forney Jr. MMSE Decision-Feedback Equalizers and Coding – Part I: Equalization Results. *IEEE Transactions on Communications*, COM-43:2582–2594, October 1995.
- [45] S. Haykin. *Adaptive Filter Theory*. Prentice-Hall, Upper Saddle River, New Jersey, Third Edition, 1996.
- [46] W. Gerstacker and J. Huber. Improved Equalization for GSM Mobile Communications. In *Proceedings of International Conference on Telecommunication*, Istanbul, April 1996.
- [47] G.D. Forney, Jr. The Viterbi Algorithm. *IEEE Proceedings*, 61:268–278, 1973.
- [48] W. Van Etten. Maximum-Likelihood Receiver for Multiple Channel Transmission Systems. *IEEE Transactions on Communications*, COM-76:276–283, February 1976.
- [49] N. Al-Dhahir. FIR Channel-Shortening Equalizers for MIMO ISI Channels. *IEEE Transactions on Communications*, COM-49:213–218, February 2001.
- [50] R. Schober and W. H. Gerstacker. Adaptive Noncoherent DFE for MDPSK Signals Transmitted over ISI Channels. *IEEE Transactions on Communications*, COM-48:1128–1140, July 2000.
- [51] R. Schober, W. H. Gerstacker, and J.B. Huber. Adaptive Noncoherent Linear Minimum ISI Equalization for MDPSK and MDAPSX Signals. *IEEE Transactions on Signal Processing*, SP-49:2018–2030, September 2001.

This page intentionally left blank

Chapter 22

FAST ROUTING AND RECOVERY PROTOCOLS IN HYBRID AD-HOC CELLULAR NETWORKS

Mostafa Bassiouni, Wei Cui, and Bin Zhou

*School of Computer Science
University of Central Florida
Computer Science Building
Orlando, Florida 32816*

{bassi, willy, bzhou}@cs.ucf.edu

Abstract A hybrid ad-hoc cellular network (HACN) is based on the idea of replacing the stationary cellular base stations with mobile routers. Communications among the mobile routers is achieved using satellite links or high bandwidth wireless channels. In this chapter, we present and evaluate efficient recovery and routing protocols for mobile routers in HACN. In the dual backup recovery protocol, a standby router is mapped (dedicated) to each primary router. A more flexible protocol, called the distributed recovery protocol, is obtained by relaxing this one to one mapping and scattering the backup routers geographically among the primary routers. The effectiveness of the distributed recovery protocol is demonstrated by simulation results. A simple analytical model is also presented for computing the blocking probability of new calls in the presence of router failures. The chapter is concluded by presenting an efficient location-based routing protocol for HACN mobile routers. Performance results of the routing protocol are presented.

Keywords: Cellular networks, ad-hoc wireless networks, fault tolerance, recovery protocols, location based routing protocols, terminal mobility.

1. Introduction

The hybrid ad-hoc cellular network (HACN) is a hybrid mobile wireless architecture that combines the advantages of both the ad-hoc and the cellular models. This hybrid network provides rapidly deployable communications capability and communications-on-the-move services with enhanced flexibility and scalability. In HACN, specialized mobile

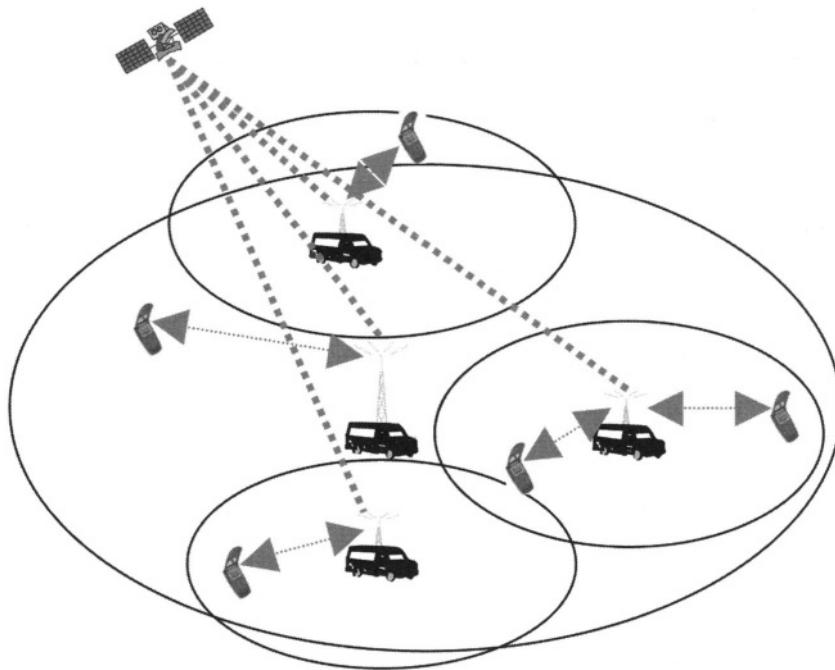


Figure 22.1. Example of a two-tier HACN

routers are used to achieve continued connectivity and fast forwarding. A mobile router has functionality similar to a cellular base station, but has no wired connections. In its simplest form, the mobile router could be a truck-mounted transceiver box with rechargeable battery. Satellite links or backbone wireless channels are used for communications among the mobile routers while short-hop wireless channels are used for communications between each mobile router and its users (mobile hosts). When the mobile hosts move, the mobile routers would also move to ensure the continuity of coverage and improve the quality of service for active connections. In one-tier (flat) HACN, the coverage areas (cells) of different mobile routers are mostly non-overlapping. A hierarchical version of HACN is obtained by adding another tier of mobile routers, called umbrella routers. An umbrella router has a more powerful transmitter and its coverage area usually overlaps that of several one-tier routers. Conceptually, a two-tier HACN is similar to the hierarchical macro-micro architecture used in cellular networks. Figure 22.1 1 shows a two-tier HACN with one umbrella cell and three regular mobile routers.

The performance of one-tier HACN and its two-tier hierarchical counterpart has been investigated in [3]. In general, the two-tier model has the potential to give better performance and better geographical coverage than the one-tier counterpart. This is especially true when the degree of the randomness of the movement of mobile terminals is high. However, under the constraint of equal power consumption and for mobility patterns with low randomness, the one tier system could provide higher throughput.

2. Recovery Protocols for HACN Mobile Routers

To improve the reliability and resiliency of HACN, mechanisms must be put in place to ensure that the operations of the mobile routers can survive environmental hazards or hostile attacks. A mobile router can become immobilized due to a flat tire, failed automotive engine, or some type of road obstruction. Although the mobility of the router is compromised, the wireless transceiver in this case is intact and can continue to provide service in a stationary mode. A more serious scenario is the failure or total destruction of the router's transceiver. This condition forces the termination of all active connections served by the failed router. In order to be able to handle these faults when they occur, the network must be equipped with robust fault tolerance and recovery protocols. These protocols normally entail adding redundant hardware and incurring some extra overhead during normal operations. Understandably, there is a tradeoff between the cost of the extra hardware and the level of robustness and gracefulness by which the faults can be handled. Below we elaborate on a new recovery protocol, called the distributed recovery protocol, that we have designed and evaluated.

In the dual backup restoration protocol [2], an extra backup router is assigned to each mobile router. The backup router exchanges state-update messages with the primary router in order to be able to accurately mirror its state. The backup router also follows the movement of its primary as close as possible. If the primary router malfunctions, its backup can take over promptly and re-establish links to the hosts served by the failed primary.

The dual backup protocol is simple and provides definite performance gains in face of router failures and hostile attacks. In particular, the dual backup protocol provides the fastest recovery when a backup router survives the destruction of its primary router. However, one disadvantage of the dual backup protocol is that the primary router and its backup are likely to be destroyed together (e.g., they enter a mine field together).

Furthermore, the dual backup protocol does not provide optimal utilization of the backup resources in the long run. As more primary routers get knocked out and are replaced by their backups, several cells will operate without backups, which ultimately degrades the gracefulness of future recovery.

Rather than strictly dedicating a backup router to each primary mobile router, the backup routers are carefully scattered among the primary mobile routers. When a primary router fails, one or more of the nearby backup routers is dispatched to provide coverage for the set of mobile terminals originally covered by the failed router. This protocol is called the distributed recovery protocol. It is flexible and does not require dedicating a backup router for each primary router.

The evaluation of the recovery protocols for HACN has been primarily done by extensive simulation tests. An analytical model has also been developed for certain useful cases. The model assumes that the time to dispatch a backup router to the location of the failed router is small compared to the time needed to repair the failed router as well as the mean time between failures. Inter-cell traffic is assumed to be of low volume, i.e., new call requests represent the dominant source of requests for mobile routers. It is possible to extend the model in order to relax most or all of the above assumptions, but we will focus in this chapter on the basic model without the added complexity of removing these assumptions.

2.1 Analysis of the Distributed Recovery Protocol

Assume that the time between failures (TBF) of mobile routers is exponentially distributed with a mean ω_p for a primary router and ω_s for a standby (backup) router. The repair time of a failed router is exponentially distributed with mean \mathfrak{R} . The interarrival times of new calls to a primary router are exponentially distributed with mean λ and the duration of each call is exponentially distributed with mean μ . Let M be the number of cells served by the HACN network and let N be the initial number of mobile routers deployed in the network, where $M \leq N \leq 2M$. Thus initially M routers are dispatched as primary routers and $N - M$ routers are used as standby routers.

Let j represent the total number of operational routers (both primary and standby) in the HACN network. The system can be represented by a birth-death Markov process where the transition rate from state j to state $j + 1$ is given by

$$\alpha_j = (N - j) * \mathfrak{R} \quad 0 \leq j \leq N$$

and the transition rate from state j to state $j - 1$ is given by

$$\begin{aligned}\gamma_j &= j * \omega_p && \text{for } 0 \leq j \leq M \\ \gamma_j &= M * \omega_p + (j - M) * \omega_s && \text{for } M + 1 \leq j \leq N\end{aligned}$$

The stationary distribution π_j of the occupancy of state j can be obtained by solving the above balance of flow equations. The blocking rate of new calls in state j is given by

$$\begin{aligned}\Phi_j &= 1 - (j/M) + j * B/M && \text{for } 0 \leq j \leq M - 1 \\ \Phi_j &= B && \text{for } M \leq j \leq N\end{aligned}$$

where B is calculated using the well-known Erlang-B formula

$$B = \frac{\frac{(\lambda/\mu)^c}{c!}}{\sum_{j=1}^c \frac{(\lambda/\mu)^j}{j!}}$$

In the above equation, c is the maximum link capacity of the mobile router. The steady state blocking rate of new calls in the HACN network can be computed as follows

$$\Phi = \sum_{j=0}^N \pi_j \Phi_j$$

The above analysis enables us to evaluate the performance of distributed recovery in a hybrid mobile network that initially has M primary routers and $N - M$ backup routers.

2.2 Performance Results of the Recovery Protocols

A detailed simulation model written in C++ was used in the evaluation of the recovery protocols for HACN. In the performance results reported in this chapter, the simulation tests used a one-tier network with 36 active mobile routers serving 1800 mobile terminals. The number of backup routers was zero for the simple (no backup) protocol, 36 for the dual backup recovery protocol, and either 18 or 36 for the distributed recovery protocol. Each mobile router has the capacity to serve 20 calls simultaneously. Call durations were exponentially distributed with mean of 180 seconds. The mean times between failures (MTBF) and repair times were also exponentially distributed with a wide range of values.

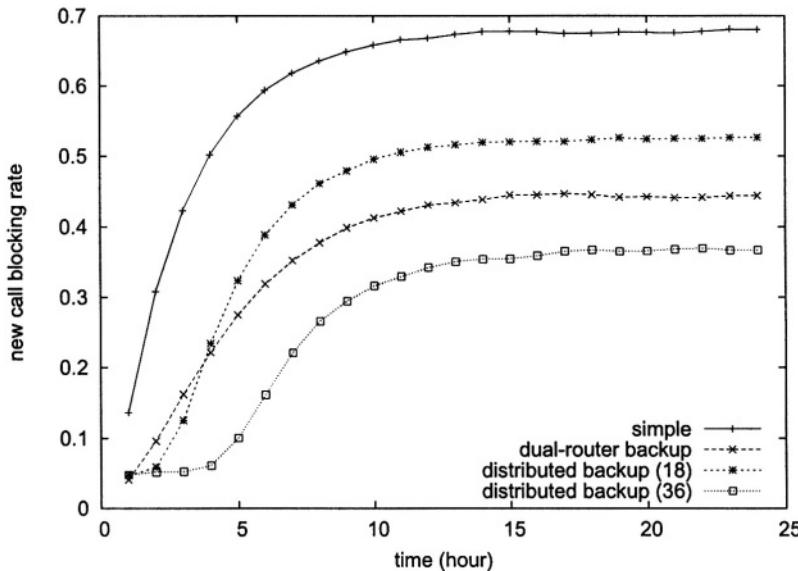


Figure 22.2. New call blocking rates for different recovery options

Figure 22.2 shows the new call blocking rate versus time for the different recovery options. In all options (including the simple protocol that does not use backups), the failed router is brought up back to operational status after a repair period. Figure 22.2 shows that the distributed recovery protocol (with 36 backup routers) outperforms the dual backup protocol which uses the same number of backup routers.

Figure 22.3 shows the new call blocking rate for the different protocols as the mean time between failures is increased from 2 to 16 hours. The results from the analytical model closely agree with those obtained by simulation. Both results confirm that the distributed protocol is quite effective and outperforms the dual router protocol under equal number of backups.

3. Location-based Routing for HACN Mobile Routers

Although HACN is a hybrid ad-hoc cellular architecture, routing messages from one mobile router to the other is basically a multi-hop ad-hoc routing problem. In this section, we present a new location-based algorithm for mobile routers different from those previously proposed in the literature [6, 8, 9].

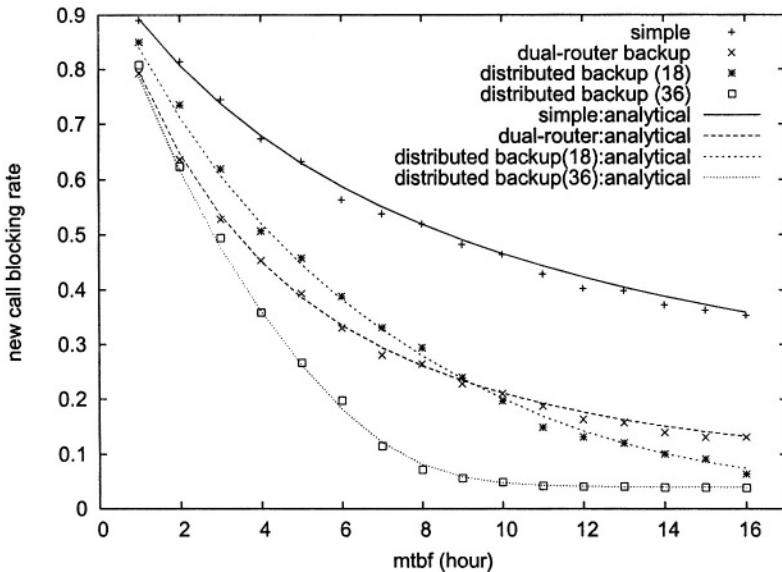


Figure 22.3. New call blocking rates vs MTBF

Location-based services (LBS) is becoming increasingly popular. New location positioning products continue to be deployed for a variety of applications with improved reliability and lower cost. Mobile phones with location positioning capabilities are rapidly spreading around the globe and it is projected that a large majority of digital wireless phones will have some position location capability in the near future. In [1], we used GPS positioning services in the design of a novel handoff prioritization technique for cellular networks and in [12] we presented a position-based multicasting (i.e., geocasting) algorithm for mobile wireless networks. In this chapter, we use similar positioning information services to improve the reachability [11] of routing algorithms in HACN. The new algorithm proposed in this chapter assumes that every mobile router can determine its position information with reasonable accuracy. This information can be easily obtained by GPS [10]. As in [7], we also assume that the source mobile router knows the approximate position of the destination mobile router. This is an acceptable assumption in the HACN environment since the number of mobile routers is an order of magnitude smaller than the number of mobile terminals and a geographical location service (GLS) that provides a translation from address to geographical location of the mobile routers can be cost-effective.

In our protocol, every mobile router periodically sends summary information about its status to its neighbors. By using the exchanged information, our algorithm seeks to reduce the number of intermediate hops while improving the chances of successful route discovery. At a given moment, the HACN network is modeled as a graph $G = (V, E)$, where V is the set of vertices (mobile routers) and E is the set of edges. An edge between one mobile router and the other means that the two routers are neighbors, i.e., they are within the transmission range of each other. The flooding algorithm is a well known routing protocol in which the routing message is sent to all neighbors and the process is repeated until the destination is reached or the algorithm fails. The flooding protocol gives the best outcome for routing success (i.e., the best reachability) since it guarantees finding the destination if there exists a feasible path to it in the network. However, this is done at the expense of higher cost of bandwidth and power consumption [4, 5]. In our performance tests, we implemented an ideal version of the flooding protocol such that the search is immediately halted once the destination is reached. The implementation of this ideal flooding algorithm is not practically feasible, but we use it in our tests as a benchmark for evaluating our location-based routing protocol for HACN.

In our algorithm, every node sends its status only to its neighbors (the terms *mobile router* and *node* are used interchangeably in the description of the routing protocol). The status information consists of the position of the node and its degree (number of neighbors within the node's transmission range). When a search is initiated from source node s to a destination node d , each node in the established path keeps track of the previous node and the next node in the path. Backtracking to the previous node is possible when the search via the next node fails and no other unvisited neighbor exists for the current node. The algorithm given in [7] is denoted GRA (geographical routing algorithm) and can be briefly described as follows. For a source node s and a destination node d , s sends its request message to the neighbor that has the shortest distance to d . That neighbor then sends the message to its neighbor with the shortest distance to d , and so on until the destination node is reached or the algorithm is stuck at some node and cannot advance. In the latter case, a route discovery protocol is used to find a new path. Basically, a depth first search algorithm is used to find an acyclic path and a routing table is needed to save the routing information.

Below we give the definitions and the notation used in our routing algorithm.

Useful degree (\ddot{U}_v) of a neighbor: For a current node c arid its neighbor v , the useful degree of v is a measure of the number of nodes

reachable from v which can be useful in continuing the search for the destination d . In the tests reported in this chapter, the useful degree of v is equal to half the number of neighbors of v , i.e., half the number of mobile routers that are within the transmission range of v . The rationale of this method is that if the nodes are randomly located, then half of the neighbors of v will be closer than v to the destination d .

Deviation angle (α_v): Consider a current mobile router c , a destination mobile router d and a neighboring mobile router of c , say v . The deviation angle of v , α_v , is defined as the angle between vector cd and cv .

Direct distance (\check{D}_v): Consider a current mobile router c , a destination mobile router d and a neighbor router of c , say v . The direct distance of v with respect to c and d is the length of the projection of vector cv on vector cd .

Routing weight (ω_v): For a current mobile router c , one of its neighbor mobile routers v and a destination mobile router d , the routing weight of v , ω_v , is defined as:

$$\omega_v = \check{D}_v * \delta(\check{U}_v)/L + \check{U}_v + f(\alpha_v) \quad (22.1)$$

Where,

$$\delta(X) = \begin{cases} 1, & X \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

Equation (22.1) has three primary parameters: the direct distance, the useful degree and the deviation angle of node v . The parameter L is tunable and is used to scale the contribution of the direct distance with respect to the useful degree. We refer to parameter L as the “level selector” (or simply the level) since it determines the level of contribution of the direct distance \check{D} on the routing weight. Smaller values of L produce higher weights for \check{D} in the selection of the next hop. The function f in equation (22.1) is the angle weight function and is used to scale the contribution of the deviation angle. This angle could have a value between 0 and π (measured symmetrically on both sides of the line cd). The contribution of the deviation angle to the routing weight of a node should decrease when the angle becomes larger. This is because a larger deviation angle represents a less favorable choice with respect to reaching the destination d .

In our simulation tests, we evaluated several choices of the value of the level parameter L and the angle weight function f . In general, we have found that the performance is improved when the direct distance \check{D} and the useful degree \check{U} have the same effective weight in equation

(22.1). The tests have shown that a good choice of the angle weight function f is a stair-wise function yielding a maximum weight that is approximately equal to the average value of \bar{U} (this maximum weight is obtained when α_v approaches 0).

3.1 Performance Results of the Routing Protocol

In our simulation tests, we used a model similar to that described in [5]. We assume the mobile routers are distributed randomly in a 1000×1000 unit area. The number of mobile routers in our tests ranged from 20 to 70. The value of the transmission radius R ranged from 200 to 500 and the level parameter L ranged from 1 to 100. For each test situation, we generated 10000 randomly distributed scenarios to get the reported results.

Reachability metric: For every test situation, we compute the fraction of scenarios in which our algorithm fails to find a route to the destination but the flooding algorithm succeeds to do so. We call this metric the reachability difference and is computed as follows: Let $Total_L$ be the total number of tests for a given setting and let $Diff_L$ be the difference between the number of cases of failure of our algorithm and that of the ideal flooding algorithm, then the reachability difference of our algorithm is defined as:

$$\text{reachability_difference} = \frac{Diff_L}{Total_L} \times 100\%$$

Another metric shown in our graphs is the number of search nodes. A search node is any node that has been visited by the search algorithm during the search process. The number of search nodes reflects the total cost (bandwidth, computation and power consumption) of the search process at all nodes. The flooding algorithm has a high number of search nodes since the search request propagates concurrently to all neighbors.

Figure 22.4 gives comparisons of our protocol, denoted HACN, with the GRA algorithm presented in [7]. The value of L was set to 60 and the transmission range was set to 200. The figure shows that the reachability of our protocol is better than that of GRA. Figure 22.5 shows that the improvement in reachability is achieved with a very small increase in the number of search nodes. The number of search nodes for HACN is almost the same as that of GRA, and they are both much smaller than the corresponding number of the ideal flooding method. The proposed protocol significantly improves the reachability of the routing algorithm without incurring any significant overhead. We are currently investigating ways to improve the selection of the algorithm's parameters in order to further improve its performance.

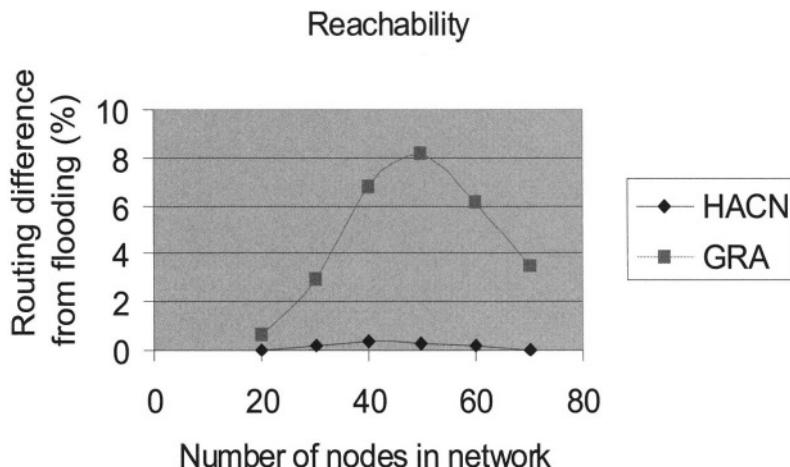


Figure 22.4. Comparison of routing reachability

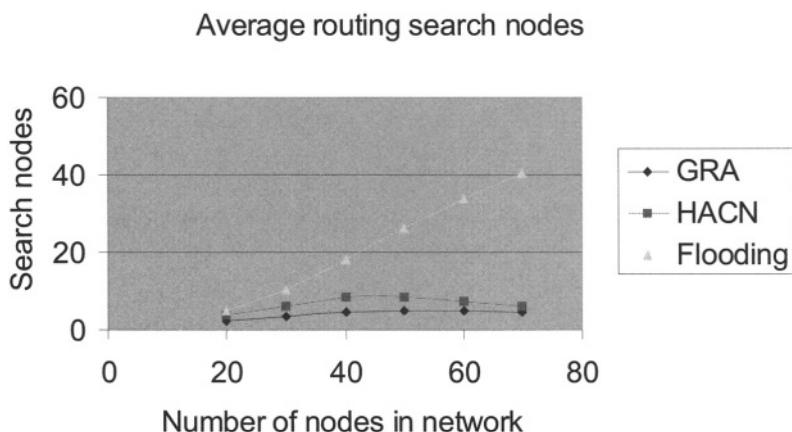


Figure 22.5. Comparison of search overhead

4. Conclusions

In this chapter, we presented a distributed recovery protocol for the mobile routers in hybrid ad-hoc cellular networks. We also presented a location based routing algorithm for these mobile routers. Simulation tests showed that both protocols are quite effective. Further research is planned for improving and fine tuning both algorithms.

Acknowledgment

This work has been supported by ARO under Grant Number DAAD19-01-1-0502 and by NSF under Grant Number 0086251.

References

- [1] M. Chiu and M. Bassiouni, "Predictive Schemes for Handoff Prioritization in Cellular Networks Based on Mobile Positioning," *IEEE Journal on Selected Areas in Communications*, vol.18, no.3, pp.510-522, March 2000.
- [2] W. Cui and M. Bassiouni, "Channel Planning and Fault Recovery in Hierarchical Hybrid cellular Networks with Mobile Routers," *Proceedings of IEEE Wireless Local Networks – 26th LCN Conference*, pp.646-652, November 2001.
- [3] W. Cui and M. Bassiouni "Analysis of Hierarchical Cellular Networks with Mobile Base Stations," *Journal of Wireless Communications and Mobile Computing*, John Wiley & Sons Publishing, vol.2, pp.131-149, March 2002.
- [4] Min-Te Sun, Wuchi Feng and Ten-Hwang Lai, "Location aided broadcast in wireless ad hoc networks," *IEEE Global Telecommunications Conference (GLOBECOM)*, vol.5, pp.2842-2846, 2001.
- [5] Young-Bae Ko and Nitin H. Vaidya, "Location-aided routing (LAR) in mobile ad hoc networks," *Wireless Networks*, vol.6, Issue4, pp.307-321, July 2000.
- [6] V. Rodoplu and T. H. Meng, "Position based CDMA with multiuser detection (P- CDMA/MUD) for wireless ad hoc networks," *IEEE Sixth International Symposium on Spread Spectrum Techniques and Applications*, vol.1, pp.336 -340, 2000.
- [7] R. Jain and A. Puri and R. Sengupta, "Geographical routing using partial information for wireless ad hoc networks," *IEEE Personal Communications*, vol.8, no.1, pp.48-57, Feb. 2001.

- [8] Yu-Liang Chang and Ching-Chi Hsu, "Routing in wireless/mobile ad-hoc networks via dynamic group construction," *Mobile Networks and Applications*, vol.5, no.1, pp.27-37, March 2000.
- [9] Yu-Chee Tseng, Sze-Yao Ni, Yuh-Shyan Chen and Jang-Pang Sheu, "The broadcast storm problem in a mobile ad hoc network," *Wireless Networks*, vol.8, Issue2/3, pp.153-167, March 2002.
- [10] G. Dommetty and R. Jain, "Potential networking applications of global positioning systems (GPS)," *Technical Report TR-24*, The Ohio State University, 1996, <http://www.cis.ohio-state.edu/~jain/papers/gps.htm>.
- [11] C-K Toh, "Wireless ATM and ad-hoc networks," Kluwer academic publishers, ISBN 079239822-x, 1997.
- [12] H. El-Aarag and M. Bassiouni, "A Reliable Congestion Control Mechanism for Geocasting in Mobile Wireless Networks," *International Journal of Network Management*, John Wiley & Sons Publishing, vol.13, no.5, pp.375-387, September 2003.

This page intentionally left blank

INDEX

- Access-based clustering protocol (ABCP), 394, 409–411, 412
- Accounting, 533
- A-Cell relay. *See* Ad hoc-Cellular relay
- ACF. *See* Auto-correlation function
- ACK-Fragmentation Protocol, 277–278
- ACM. *See* Adaptive modulation / coding
- Active location update, 477
- Active mode in mobile hosts, 519
- Adaptation paths, 521–522
- Adaptive modulation / coding (ACM), 158, 174–175, 672, 673
- Additive white Gaussian noise (AWGN), 84, 99, 104, 116, 125–126, 136, 180
- diagonal STBCs and, 658
 - dual-hop systems and, 449, 456
 - multi-hop systems and, 447
 - multiple antennas and, 618, 630, 633–634, 635, 638
- Addressing, 2, 11–12
- Ad hoc-Cellular (A-Cell) relay, 53–66
- architecture of, 62–63
 - modeling of, 64–66
 - related work, 55–60
- Ad hoc networks, 383–423. *See also*
- Clustering; Hybrid ad-hoc cellular networks
 - security in, 534, 536–542
 - sensor networks compared with, 29
- Ad hoc relay stations (ARSs), 56
- Adjustment loop, 157, 170–171
- Advanced Mobile Videophone, 18
- Agent Management System (AMS), 543–544
- Alamouti's code, 639–643, 656, 657, 661, 674, 675, 676–677
- Alteration of information, 532
- Alternative location algorithm (ALA), 477
- Amount of fading (AF), 445
- Amplify-and-forward relaying, 445
- Analog relaying, 445
- Anomaly detection, 535
- Antenna-array techniques, 83, 106–107
- Antennas
- directive, 53, 54, 62–63, 66
 - multiple (*see* Multiple antennas)
 - smart, 78, 80, 106, 140–141
- AOD. *See* Average outage duration
- Aperiodic auto-correlation function (ACF), 89, 99, 101
- Aperiodic cross-correlation function (CCF), 89, 99, 112, 129
- Application Program Interface (API), 9, 12, 362, 364, 380
- ARAN, 537
- Area residence time (ART), 486–488
- ARIADNE, 539
- ARIB, 84, 93
- ARPANET, 12
- Asymmetric cryptography, 533, 546–547
- Asymptotic stability, of discrete-time linear systems, 189–190
- Asynchronous invalidation, 518–519
- Authentication, 534
- biometric, 533
 - data, 547, 548
 - IPSec, 340, 344, 345, 348–353, 354–355
 - purpose of, 532
- Auto-correlation function (ACF), 83, 90t, 109, 139
- aperiodic, 89, 99, 101
 - characteristics of, 89
 - of complementary codes, 94, 96, 97f
 - in-phase, 89
 - near-far effect and, 102
 - out-of-phase (*see* Out-of-phase auto-correlation function)
 - of OVSF codes, 93, 94
- Auto-interference, 182, 193
- Autonomous registration, 477
- Average bit error rate (BER), 444, 454–457, 466–467
- Average outage duration (AOD), 209–236
- method of analysis, 212–214
 - system model, 218–219

- AWGN. *See* Additive white Gaussian noise
- Background noise, 430, 434
- Back-to-back sinusoidal waveform, 586
- Band-limited nonlinear channels, 589–599
- Bandwidth efficiency, 559, 560–563, 565–567, 568–569, 573–574, 580, 581
comparison by modulation, 586–587
in nonlinear channels, 589–599
- Base-Centric Routing (BCR), 58–59
- Base stations (BS)
antenna-array techniques and, 106–107
CDMA, 88, 158–159, 175
adjustment loop and, 170–171
reduced active set and, 172–173
soft handoff and, 157, 165–167
SSDT and, 171–172
TPC rate reduction and, 168–169
TPC synchronization with, 168
dual-hop systems and, 453, 457
- Internet link enhancements and, 295
mobility management and, 474, 475f
security and, 532–533, 548–549
sensor networks and, 548–549
soft handoff and, 199
TCP and, 246, 248, 280–283
uplink load and, 426–427
up-link synchronization and, 109
- Base transceiver stations (BTS), 321–324, 329
- BCR. *See* Base-Centric Routing
- Beacons, 391
- Beamforming, antenna, 78, 80, 106–107, 140–141
- Behavior-based intrusion detection, 535
- Bent codes, 91
- BER. *See* Bit error rate
- Berkeley Snoop, 296, 298, 299, 305–306
- Best S+I algorithm, 219, 220, 224–227, 236
- Best signal power algorithm, 220, 221–222, 236
- Best SIR algorithm, 220, 223–224, 227, 236
- Beyond 3G (B3G) wireless communications, 84, 85
- Binary phase shift keying (BPSK), 131, 136
diagonal STBCs and, 664, 665, 676, 677
pulse shaping waveforms and, 564, 566, 567
- Biometric authentication, 533
- Birth-death Markov process, 46, 688–689
- Bit error rate (BER). *See also* Average bit error rate
of CC-CDMA, 116–117, 120, 125, 126, 127f
of CC/DS-CDMA, 133–134, 135, 136
diagonal STBCs and, 661, 665
Eb/No compared with, 161f
for MLSD, 664, 666f, 667f, 668f, 675–676
- Bit error rate (BER) (*continued*)
outer loop power control and, 200
pulse shaping waveforms and, 568, 599–604
of third generation CDMA, 161f, 162, 163f, 165, 166
TPC commands and, 162, 163f, 166
- BLAST interface, 645–647
- BLER. *See* Block error rate
- Blind relays, 445, 455
- Block error rate (BLER), 158, 165, 168, 175, 200, 327
- Block-fading (BF) channel, 617, 628
- BPSK. *See* Binary phase shift keying
- Busy-line effect, 478, 486–487, 490
- Cache coherence, 518–519
- Cache consistency, 526–527
- Cache replacement policies, 519–520
- Caching, 513, 514, 517–528
architecture for, 523–528
current mechanisms, 517–520
extending current mechanisms, 520–522
hierarchical, 521–522
multimedia delivery, 527–528
optimal set of documents, 526
role in pervasive computing, 522–523
utility functions, 524–526
- Call blocking, A-Cell, 66
- Call delivery. *See* Location tracking
- Call Server, 363, 364, 367–368, 369–379, 380
- Call setup, 369–373
- Call state control function (CSCF), 13, 360
- Call-to-mobility ratio (CMR), 490, 498
- Call transfer, 374–376
- Carrier-to interference (C/I) ratio, 326, 327
- Carrier-to-total-interference ratio, 428–429, 430, 432, 433, 434–435, 437–439
- CATT, 84, 93
- CC-Code Division Multiple Access (CDMA), 83, 85, 94, 114–129, 130–131
in MAI-free channels, 114–120
in multipath channels, 120–127
pros and cons for, 127–129
- CC/DS-Code Division Multiple Access (CDMA), 84, 85, 142
merits of, 139–141
MI-free IST and, 129–138
- CCF. *See* Cross-correlation function
- CCI. *See* Co-channel interference
- CCL. *See* Computer and Communication Research Laboratory
- CDMA. *See* Code Division Multiple Access
- Cell boundary crossings, 486–490
- Cell residence time (CRT), 486, 487, 492, 493, 495
- Cellular phones. *See* Mobile phones
- Certificate authority (CA), 533–534, 537

- Certificate chains, 538–539
Certificate graphs, 538
Channel capacity, 615, 616–634
 - with delay constraints, 630–631
 - deterministic channel and, 617–618
 - ergodic Rayleigh fading channel and, 619–626, 631
 - with no delay constraints, 629–630
 - nonergodic Rayleigh fading channel and, 626–631
Channel impulse response (CIR), 658–659, 660, 661, 662, 664, 665, 669, 675
Channel state information (CSI), 617, 619, 629, 631–634
 - imperfect, at receiver, 633
 - none, 633
 - perfect, at receiver, 633, 635, 640
 - space-time codes and, 643–644
 - at transmitter and receiver, 633–634
Channel state information (CSI)-assisted relays, 445, 448, 449–450, 452, 454
CHAP, 338
Charging applications, 13
Cholesky factorization, 670, 671
CIR. *See* Channel impulse response
Cisco Systems, Inc., 335, 339
Closed-loop poles, 189
Closed-loop power control, 83, 87, 102, 140, 158–159, 161, 186
Closed-loop sector of up-link synchronization, 108, 109, 110f
Closed-loop transfer function of the local loop, 196–197
Closed loop transmit diversity, 74–75
Clusterheads, 386, 395, 396, 397, 401, 402, 404, 405, 409–410
 - load distribution, 394
 - number of changes in, 393
 - power control and, 411
Clustering, 383–423
 - algorithms, 394–412
 - LEACH, 32–33, 49
 - performance metrics, 392–394
 - protocol performance and, 387–390
 - radio resource management and, 387–390
 - system model, 391–392
 - transmission range impact, 413–416
 - transport protocol and, 390
Cluster member power control, 411
Cluster membership changes, 393, 404
CMR. *See* Call-to-mobility ratio
Co-channel interference (CCI), 106–107, 209, 210
 - impact on LCR and AOD of MRC, 211–217, 235–236
 - impact on LCR and AOD of SC, 218–234
Code Division Multiple Access (CDMA), 3, 60, 83–149, 425–440. *See also* CC-Code Division Multiple Access; Direct sequence Code Division Multiple Access; Isotropic CDMA air-link technologies; Third generation Code Division Multiple Access; Time Division Synchronous Code Division Multiple Access; Uplink load; Wideband Code Division Multiple Access
 - basics of codes, 88–97
 - C/I ratio in, 327
 - frequency hopping, 85, 87
 - pilot-aided, 83, 107–108
 - time hopping, 84, 85, 87
 - traditional codes, 90–97
Code Division Multiple Access-2000 (CDMA2000), 84, 93, 103, 111, 160, 173
Coding gain, 636
Common Packet Channel (CPCH), 74
Communication compromise, 532
Complementary codes (CCs), 83, 91. *See also* CC-Code Division Multiple Access
 - characteristics of, 94–97
 - with PG 8 to 12, 144–149
Compulsory tunnels, 336–337
Computer and Communication Research Laboratory (CCL), 359, 360
 - reference architecture of, 363–364, 365f, 380
CONFIDANT. *See* Cooperation of nodes and fairness in dynamic ad hoc networks
Congestion control
 - Differentiated Services and, 309, 310
 - fairness of, 246
 - in K-SACK, 271–272
 - in TCP, 243–244, 253, 267, 279, 282
 - in TCP Reno, 249
 - in TCP Santa Cruz, 274–276
 - in TCP Tahoe, 249
 - in TCP Vegas, 251–252, 258
 - in TCP Westwood, 273
 - in Wireless TCP, 278
Content adaptation, 521–522
Continuous media (CM) applications, 527–528
Continuous phase frequency shift keying (CPFSK), 569
Continuous phase modulation (CPM), 562–563
Control and signaling plane of Softswitch, 362

- Cooperation of nodes and fairness in dynamic ad hoc networks (CONFIDANT), 540
- Cooperative / collaborative diversity, 444, 458
- Cooperative games, 550–552
- Corporate Authentication Server, 319
- Correctness, of PFR, 44, 45–46
- Correlated wireless channel error, 262–268, 270, 273, 274
- Covered nodes, 402
- Coxian model, 504
- CPCH. *See* Common Packet Channel
- CPE. *See* Customer premises equipment
- CPM. *See* Continuous phase modulation
- Cross-correlation function (CCF), 83, 91t, 101, 109, 116, 141
- aperiodic, 89, 99, 112, 129
 - beamforming algorithms and, 107
 - characteristics of, 89–90
 - of complementary codes, 94, 96, 98f
 - of Gold codes, 92
 - in-phase, 112
 - of Kasami codes, 92
 - of M-sequences, 91
 - near-far effect and, 102
 - out-of-phase (*see* Out-of-phase cross correlation function)
 - of OVSF codes, 93, 94
 - RAKE and, 105
- Cross-coupling relative load, 196
- CRT. *See* Cell residence time
- Cryptography / encryption, 533, 534, 549.
- See also* Data Encryption Standard; Secure Network Encryption Protocol
 - asymmetric, 533, 546–547
 - IPSec, 340, 344, 345, 346–348, 353–355
 - Microsoft point-to-point, 338
 - SNEP, 548
 - symmetric, 340, 344, 533, 539
 - threshold, 537–538
- CSI. *See* Channel state information
- Customer premises equipment (CPE), 321–322, 328
- Customised Infotainment (CI), 15, 16
- Data authentication, 547, 548
- Data confidentiality, 547, 548
- Data distance, 520
- Data Encryption Standard (DES), 340, 344, 345, 353
- triple, 346–348, 353–355
- Data freshness, 547, 548
- Data integrity, 547, 548
- Data privacy, 532–533
- Data staging, 522
- DCH. *See* Dedicated Channel
- Decision-feedback equalization (DFE), 655, 656, 657, 660, 669–671, 672–674, 676–677
- Decision-feedback sequence estimation (DFSE), 655, 671–672, 673, 675, 676, 677
- Decode-and-forward relaying, 445
- Decorrelating detectors, 103–104
- Dedicated Channel (DCH), 73, 75, 76, 77
- Delay-limited capacity, 617
- Denial of service attacks, 532, 536, 541–542, 547
- DES. *See* Data Encryption Standard
- Destination Sequenced Distance Vector (DSDV) protocol, 539
- Desynchronization, 548
- Detection Points (DP), 366
- Deterministic channel, 617–618
- DFE. *See* Decision-feedback equalization
- DFSE. *See* Decision-feedback sequence estimation
- DHCP. *See* Dynamic Host Configuration Protocol
- D-hop clusters, 394, 399
- Diagonal space-time block codes (STBCs), 655–678
- complexity issues, 673–675
 - design criterion for, 659–660
 - design of, 661–666
 - equalization and, 667–675
 - notations for, 657–658
 - simulation results, 675–677
 - transmission model for, 658–659
- Differential phase shift keying (DPSK), 455, 456, 457
- Differential unitary space-time signals, 644
- Differentiated Services (DS), 301–302, 304, 307, 308–311, 314
- Diffie-Hellman key agreement, 549
- Digital relaying, 445
- Digital signature based routing, 537–539
- Digital Subscriber Line (DSL) modems, 77
- Diophantine equation, 201, 202
- Directed diffusion, 30–32, 42
- Directive antennas, 53, 54, 62–63, 66
- Directory Facilitatory (DF), 543–544
- Directory registers, 477
- Direct sequence Code Division Multiple Access (DS-CDMA), 85, 87–88, 134, 136–138, 139, 142, 199. *See also* CC / DS-Code Division Multiple Access
- Direct transmission, 445–448, 461–462
- Disclosure of information, 532
- Discouraged arrivals Markovian process, 46
- Discrete-time linear systems, asymptotic stability of, 189–190
- Dispatch service, 17–18

- Distance-based location update, 477–478, 480
Distributed power control (DPC), 185
Distributed recovery protocol, 685, 687, 688–690
Distributed trust, 543–544
Diversity
 average BER and, 457
 closed loop transmit, 74–75
 cooperative / collaborative, 444, 458
 multipath, 139
 open loop transmit, 74
 optimum power allocation and, 459–460
 outage probability and, 453–454
 spatial, 57–58
 systems without, 458–459
Diversity gain, 167, 168, 217, 636
Downlink power control, 165
 properties of, 200–201
 during soft handoff, 166–171
Downlink Shared Channel (DSCH), 73
Downlink signal-to-interference ratio (SIR), 183, 200–201
DPSK. *See* Differential phase shift keying
DS-CDMA. *See* Direct-sequence Code Division Multiple Access
DSCH. *See* Downlink Shared Channel
DSDV protocol. *See* Destination Sequenced Distance Vector protocol
DSL modems. *See* Digital Subscriber Line modems
Dual backup recovery protocol, 685, 687–688, 689–690
Dual half sinusoidal waveform, 586
Dual-hop systems, 448–464, 470
 optimum power allocation in, 458–464
 outage probability in, 450–454, 458, 460
 system and channel models, 448–450
Dynamic Host Configuration Protocol (DHCP), 324
Dynamic Source Routing, 539
EAD, 48–49
Eavesdropping, 547
Eb / No. *See* Energy per bit to noise ratio
EDGE. *See* Enhanced Data Rates for Global Evolution
Eifel algorithm, 279–280
Encryption. *See* Cryptography / encryption
Encryption client, 324–325
Energy balance, 47–48
Energy consumption
 LEACH Protocol and, 32–33
 STEM Protocol and, 33–36
 TCP and, 254, 259, 260, 261f, 267
Energy efficiency
 of PFR, 44, 46, 48
 of sensor networks, 29
Energy per bit to noise ratio (Eb/No), 158, 163, 164, 173–175
 BER compared with, 161f
 estimating, 164
 inner loop power control and, 159
 outer loop power control and, 165
 portability architecture and, 327
Enhanced Data Rates for Global Evolution (EDGE), 4, 75, 664
Enhanced Messaging Services (EMS), 11
Envelope fluctuation factor, 587–589, 605
Equal gain combining (EGC)- RAKE, 105, 139–140
Ergodic Rayleigh fading channel, 619–626, 631
Erlang-B formula, 689
Euclidean distance, 637–638, 662, 663
European Telecommunications Standards Institute (ETSI), 57, 84, 93
Explicit Congestion Notification (ECN), 279
Exponentially weighted moving average (EWMA) scheme, 519
Exposed node problem, 387–388
Fabrication, 532
FACH. *See* Forward Access Channel
FACK Transmission Control Protocol (TCP), 241, 246
 compared with other variants, 253
 dynamics and performance in, 254–262, 264–269
 maximum throughput, 269–270
 properties of, 251
Fading
 amount of, 445
 dual-hop systems and, 449–450, 452, 454–455
 impact on LCR and AOD of MRC, 211–217
 impact on LCR and AOD of SC, 227–234
 multipath, 158, 160–161, 254
 Nakagami (*see* Nakagami fading)
 Rayleigh (*see* Rayleigh fading)
 Rician (*see* Rician fading)
Fading inter-symbol interference (ISI)
 channels, 655–678. *See also* Diagonal space-time block codes
Fault tolerance, 30, 48, 687
FDD. *See* Frequency Division Duplex
FDMA. *See* Frequency Domain Multiple Access
Feasibility relative load, 194
FEC. *See* Forward Error Correction
Feedback (FB) filters, 669, 673–674
Feedforward (FF) filters, 669, 671, 672, 673–674
FH-CDMA. *See* Frequency hopping Code Division Multiple Access

- Fingers, RAKE receiver, 104–105, 120, 130, 133, 139
- Finite impulse response (FIR) minimum mean-squared error (MMSE), 669, 671
- Firewalls, 536, 537
- First-party call control, 368
- Fixed gain relays, 452–453
- Flooding, 536, 540, 541
- clustering and, 389, 399–400
 - hybrid ad-hoc cellular networks and, 692, 694
 - sensor networks and, 31, 32
- Forward Access Channel (FACH), 73, 74
- Forward Error Correction (FEC), 296, 300, 301, 302, 327
- Freeze-TCP, 277, 294
- Frequency Division Duplex (FDD), 161
- Frequency Domain Multiple Access (FDMA), 86, 104
- Frequency-flat Rayleigh fading, 616–617
- Frequency hopping Code Division Multiple Access (FH-CDMA), 85, 87
- Frequency shift keying (FSK), 561, 569
- Game theory, 541, 550–552
- Gateway GPRS support nodes (GGSNs), 360
- Gateway location register (GLR), 476
- Gateway nodes, 386, 395, 396, 397, 401
- Gaussian distributions, 669–670
- Gaussian interference, 180
- Gaussian minimum-shift keying (GMSK)
- diagonal STBCs and, 664
 - pulse shaping waveforms and, 561, 562–563, 564, 567, 569, 570–571, 573
 - spectral characteristics of, 571
- Gaussian noise, 602. *See also* Additive white Gaussian noise
- Gaussian power control errors, 191
- Gaussian pulse shaping filters, 571
- Gaussian random variables (RV), 634, 636, 658
- Gaussian tail functions, 627
- Gaussian waveform generators, 564, 567, 569
- General Packet Radio Service (GPRS), 4, 14, 242, 243f, 360
- Generic Routing Encapsulation (GRE), 337, 338, 339–340
- Geographical location service (GLS), 691
- Geographical routing algorithm (GRA), 692, 694
- Geographic positioning system (GPS), 25, 42, 53, 54, 62, 478, 691
- GERAN. *See* Global System for Mobile Communications Enhanced Radio Access Network
- GGSNs. *See* Gateway GPRS support nodes
- Global System for Mobile Communications (GSM), 3, 4, 6, 75, 186. *See also* Mobile Application Part from GSM
- diagonal STBCs and, 664
 - pulse shaping waveforms and, 560, 561, 563, 571
- Global System for Mobile Communications Enhanced Radio Access Network (GERAN), 14
- GLR. *See* Gateway location register
- GMSK. *See* Gaussian minimum-shift keying
- GMW codes, 91
- Go Back N, 296, 298, 299
- Golay code, 637
- Gold codes, 83, 94, 96
- CC-CDMA and, 116–117, 119
 - characteristics of, 91–92
 - DS-CDMA and, 134, 136, 138
- 3GPP. *See* Third Generation Partnership Project
- GPRS. *See* General Packet Radio Service
- GPS. *See* Geographic positioning system
- Grain particles. *See* Smart dust particles
- GRE. *See* Generic Routing Encapsulation
- Greedy-Dual caching strategies, 525
- GSM. *See* Global System for Mobile Communications
- H.323 (protocol), 362, 363, 370, 373f, 374–375
- Half-cycle sinusoidal waveforms, 564, 567, 569, 571, 572, 581, 582, 585
- Handoff, 474. *See also* Soft handoff
- Handshaking protocol, 543–544
- Hard capacity, 429–430
- Hard resources, 425–426
- HARQ. *See* Hybrid Automatic Repeat Request
- HCN. *See* Highest connectivity clustering algorithm
- Hidden node problem, 387–388
- Hierarchical caching, 521–522
- Hierarchical Threshold Sensitive Energy Efficient Network (H-TEEN), 48
- Highest connectivity (HCN) clustering algorithm, 394, 405, 412
- node density and, 416–417
 - node mobility and, 417–418
 - properties of, 402–404
 - transmission range impact, 413–416
- High Speed Downlink Packet Access (HSDPA), 14, 71, 74, 79, 81, 82, 175
- described, 5, 76–78
- High Speed Downlink Shared Channel (HS-DSCH), 76–77, 200
- High Speed Shared Control Channel (HS-SCCH), 77

- Hilly terrain (HT) channel model, 661, 675, 676
HLC. *See* Home location cache
HLR. *See* Home Location Register
HMAC-MD5, 345, 349–351, 354
HMAC-SHA-1, 345, 352–353, 354
Hochwald / Sweldens codes (HS-C), 662–663, 664, 666, 673, 675, 676
Home location cache (HLC), 518–519
Home Location Register (HLR), 474, 477, 480, 481–482, 483–484, 491, 495, 499, 500–501, 503
Home Subscriber Server (HSS), 13, 360
Homing attacks, 548
Hops efficiency, 37–41
Host-based intrusion detection, 535
Host compromise, 532
HSDPA. *See* High Speed Downlink Packet Access
HS-DSCH. *See* High Speed Downlink Shared Channel
HSS. *See* Home Subscriber Server
HS-SCCH. *See* High Speed Shared Control Channel
H-TEEN. *See* Hierarchical Threshold Sensitive Energy Efficient Network protocol
Hybrid ad-hoc cellular networks, 685–696
 location-based routing for, 690–694
 performance of recovery protocols, 689–690
 performance of routing protocol, 694
 recovery protocols for, 687–690
Hybrid Automatic Repeat Request (HARQ), 71, 76, 77, 79
Hybrid wireless network, 59–60, 61f
Hyper-Erlang model, 504–506

ICallControl, 367–368
ICANN, 12
iCAR. *See* Integrated Cellular and Ad hoc Relay
ICMP control messages, 281
Ideal end-to-end Transmission Control Protocol (**I^{E²}**-TCP), 269–270
IETF, 11, 12, 14
IMS. *See* Internet protocol Multimedia Subsystem
IMT-2000, 5, 12, 20
IMT-TC, 5
IN call model, 365–366, 367, 380
Indirect-Transmission Control Protocol (I-TCP), 280, 283
Indoor penetration, 326
Industrial Technology Research Institute (ITRI), 359, 360
Information theft, 536
In-house portability, 328
Inner loop power control, 157, 158–159

In-phase auto-correlation function (ACF), 89
In-phase cross-correlation function (CCF), 112
Integrated Cellular and Ad hoc Relay (iCAR), 56
Integrated Services, 302, 309
Inter-arrival time, 486–487, 490, 491
Intercell interference, 430–433
Inter-cell regional (Intra-RNC) portability, 329
Interception, 532
Interest messages, 31
Interference
 auto-, 182, 193
 co-channel (*see* Co-channel interference)
 Gaussian, 180
 intercell, 430–433
 inter-symbol (*see* Inter-symbol interference)
 intracell, 430–432
 multipath (*see* Multipath interference)
 multiple access (*see* Multiple access interference)
 spatial, 614
 standard function, 194–195
Interference expansion factor, 431–432
Interference power
 measurement based estimates, 432–433
 path gain measurement based estimates, 437–439
 uplink load and, 427–429, 430–433, 437–439
International Softswitch Consortium (ISC), 359
 reference architecture, 361–362, 379
Internet, 521, 535–536. *See also* Mobile Internet; Multi-service wireless Internet link enhancements; Portability architecture
Internet protocol (IP), 7–8, 319. *See also* entries beginning with IP
Internet protocol Multimedia Subsystem (IMS), 7–19, 21
 additional deployment requirements, 18–19
 addressing, 11–12
 call model, 10–11
 network description, 13–14
 service evolution with, 14–18
Internet service provider (ISP), 324, 325
Inter-path delay, 100
Inter-Radius portability, 330–332
Inter-RNC regional portability, 330
Inter-service time, 478, 479, 486–491, 492, 493, 494, 495, 497, 498–500, 501–506
Inter-softswitch call, 377–379

- Inter-symbol interference (ISI), 100, 561, 562, 571, 573, 602. *See also* Fading inter-symbol interference channels
- Interworking domain of softswitch, 362
- Intracell interference, 430–432
- Intra-RNC portability, 329
- Intrusion detection, 534–536
- Intuitive methods, 563, 569–574, 605
- Invalidation report (IR), 517–518
- Invisibility, 523, 542
- IP network, 325
- IP Pool Server, 324
- IProtocolControl, 368
- IPSec Authentication Header (AH), 340, 341–342, 345–346
- IPSec Encapsulating Security Payload (ESP), 340, 343–344, 345–348
- IPSec protocol, 344–355
- limitations of, 353–354
 - space complexity, 345–346
 - time complexity, 346–348
- IPSec tunnels, 340–344
- IP transparency, 2
- IP transport domain of softswitch, 361
- IP tunnelling. *See* Tunnelling
- IPv4, 6, 7, 12, 307, 309
- IPv6, 7, 8, 12, 18, 20, 310
- IS-41, 474, 476, 477, 481
- cost analysis for, 490–491
 - properties of, 479–480
- IS-95, 84, 86, 93, 102, 129, 165
- IS-200, 173
- ISC. *See* International Softswitch Consortium
- ISI. *See* Inter-symbol interference
- Isotropic CDMA air-link technologies, 83, 109–141. *See also* Isotropic spreading techniques
- Isotropic MAI free property, 113, 114–129, 131–133, 139, 140–141
- Isotropic MI free property, 113, 114, 129–138, 139, 140–141
- Isotropic spreading codes (ISC), 113, 129, 130
- Isotropic spreading modulation (ISM), 113, 127, 129
- Isotropic spreading techniques (IST), 142.
- See also* Isotropic MAI free property; Isotropic MI free property
 - applications for, 111–113
 - properties of, 113–114
- Iterative power control algorithm, 195
- Jammer, 541
- Java, 12, 19
- Java Virtual, 12
- Joint probability density function (JPDF), 213, 220, 221, 229
- Karn's RLP, 296–297, 298, 299, 305, 306
- Kasami codes, 83, 91, 92
- Keyed-Hashing for Message Authentication. *See* HMAC
- Key management, 549–550
- K-hop clustering, 386, 394
- Knowledge-based intrusion detection, 535
- K-SACK, 270–273
- Laguerre polynomial, 620, 622
- Laplace transforms, 488, 493, 495, 497, 500, 502, 504, 506
- Large-set Kasami codes, 92
- Layered space-time codes, 634
- Layer three tunnelling, 340–344
- Layer two forwarding (L2F), 335, 339
- Layer two tunnelling, 335–340
- Layer two tunnelling protocol (L2TP), 321, 322, 324, 325, 331, 335–337
- LCA. *See* Linked cluster algorithm
- LCC. *See* Least cluster change algorithm
- LCR. *See* Level crossing rate
- LEACH Protocol. *See* Low-Energy Adaptive Clustering Hierarchy Protocol
- Least cluster change (LCC) algorithm, 394, 412
- node density and, 416–417
 - node mobility and, 417–418
 - properties of, 404–405
 - transmission range impact, 413–416
- Least-mean-square (LMS) algorithms, 656
- Level crossing rate (LCR), 209, 210–211
- of MRC, CCI impact on, 211–217, 235–236
 - of SC, CCI impact on, 218–234
 - of SC, fading impact on, 227–234
- L2F. *See* Layer two forwarding
- LID algorithm. *See* Lowest ID algorithm
- Linear interfaces, 644–645
- Linear minimum mean-squared error (MMSE) interface, 644–645
- Link activation, 406
- Link based estimates, 436–437
- Linked cluster algorithm (LCA), 394, 395–399, 400, 415
- Link failure, 393–394, 406, 407
- Link Level Control (LLC) protocol, 242
- Link vector algorithm (LVA), 390
- Local anchor scheme, 477
- Localized scalability, 542–543
- Local Target Protocol (LTP), 36–41, 47, 48
- Location area (LA). *See* Registration area
- Location-based routing, 690–695
- Location Based Services (LBS), 11, 15, 16
- Location cache scheme, 477
- Location-dependent data service, 520
- Location management. *See* Mobility management
- Location privacy, 533
- Location tracking (call delivery), 476, 478, 484, 485–486, 498

- Location updates, 476, 478, 485, 487, 491–495, 501, 503
active, 477
average number of, 492–493
distance-based, 477–478, 480
tradeoff cost analysis, 493–495
- Log-linear algorithms, 196
- Low-Energy Adaptive Clustering Hierarchy (LEACH) Protocol, 32–33, 49
- Lowest ID (LID) algorithm, 394, 405, 411, 412
node density and, 416–417
node mobility and, 417–418
properties of, 401–402
transmission range impact, 413–416
- LTP. *See* Local Target Protocol
- L2TP. *See* Layer two tunnelling protocol
- MAC. *See* Medium Access Control; Message Authentication Code
- MAI. *See* Multiple access interference
- Management plane of Softswitch, 362
- Markovian models, 55, 65, 66
- M-ary CDMA systems, 88, 92
- M-ary differential phase shift keying (M -DPSK), 455
- M-ary phase shift keying (M -PSK), 455, 661, 662
- M-ary quadrature amplitude modulation (M -QAM), 455, 661, 662
rotated, 663–664
- Masking uneven conditioning, 543
- Masquerading, 532
- Massachusetts Institute of Technology (MIT), 25
- Maximal ratio combining (MRC), 210
CCI impact on LCR and AOD of, 211–217, 235–236
dual-hop systems and, 457
- Maximal ratio combining (MRC)-RAKE, 105, 107, 108, 130, 133, 139
- Maximum-length sequences. *See* M-sequences
- Maximum likelihood detection, 635–639, 644
- Maximum-likelihood sequence detection (MLSD), 661, 665, 666f, 667f, 668f, 672–673, 675–676, 677
- Max-min D-clustering algorithm (MMD), 394, 399–400
- MBAC. *See* Mobility-based adaptive clustering algorithm
- MCN. *See* Multi-hop Cellular Network
- MD5. *See* HMAC-MD5; Message Digest MD5
- M-DPSK. *See* M-ary differential phase shift keying
- Media gateway, 360
- Media Gateway Control Function (MGCF), 13
- Media gateway controller (MGC), 360
- Media Gateway Control Protocol (MGCP), 362, 363
- Medium Access Control (MAC), 30, 302
clustering and, 387–389
portability architecture and, 323, 326
- Megaco, 362, 363, 369–371, 373f, 374–375, 378, 379
- MEMS. *See* Micro-electro-mechanical systems
- Message Authentication Code (MAC), 548, 549
- Message Digest MD5, 340, 344, 348–349, 354
- Message injection, 547
- Message modification, 547
- Message replaying, 547
- METP. *See* Mobile-End Transport Protocol
- MGC. *See* Media gateway controller
- MGCP. *See* Media Gateway Control Protocol
- MGF. *See* Moment generating function
- MI. *See* Multipath interference
- Micro-electro-mechanical systems (MEMS), 23, 25
- Microsoft point-to-point encryption (MPPE), 338
- Microsoft Windows 95, 338
- Microsoft Windows 98, 338
- Microsoft Windows NT, 338
- Microsoft Windows XP.NET, 9
- Middleware techniques, 513, 516–517
- MIMO system. *See* Multiple-input, multiple output system
- Minimum mean-squared error (MMSE), 103, 104, 669, 671
- Minimum mean-squared error (MMSE)-BLAST interface, 645, 647
- Minimum mean-squared error (MMSE)-linear interface, 644–645
- Minimum shift keying (MSK), 561, 562, 564, 567, 586, 592, 598, 599
intuitive methods and, 569–574
piece-wise fitting and, 579
QORC compared with, 573f
spectral characteristics of, 569–570
time-domain convolution and, 582, 584
- Minimum shift keying triangular cosine (MSKTC), 590, 597–599, 600f, 601f
- Min-two Uniform Targets Protocol (M2TP), 40–41
- Misuse detection, 535
- Mixed media mobile access, 17
- MLSD. *See* Maximum-likelihood sequence detection
- MMD. *See* Max-min D-clustering algorithm
- MMDS. *See* Multichannel Multipoint Distribution Service
- MMS. *See* Multimedia Messaging Service

- MMSE. *See* Minimum mean-squared error
 Mobile Application Part from GSM (GSM-MAP), 6, 474, 479–480, 491
 Mobile-End Transport Protocol (METP), 281–282, 283
 Mobile Internet, 1, 2, 12, 15, 18–20
 Mobile networks, 1–21. *See also* Internet protocol Multimedia Subsystem; Mobile next generation networks extending standardization scope, 12 future infrastructure platforms, 6–12 new infrastructures for, 3–6 security in, 3, 8, 531–546 SIP-based, 7–9, 12, 13, 16, 17, 19
 Mobile next generation networks (NGNs) enabling technologies for, 2–3 extended standardization for, 12 targets of, 1–2
 Mobile phones, 541, 560, 691
 Mobile routers, 685–687 location-based routing for, 690–695 recovery protocols for, 687–690
 Mobile stations (MS), 453, 457, 475f
 Mobile switching centers (MSC), 320, 474, 475f, 476, 485, 491
 Mobile switching center (MSC) server, 13
 Mobile switching stations (MSS), 518
 Mobility agents (MA), 481, 482, 498, 499
 Mobility-based adaptive clustering (MBAC) algorithm, 394, 405–409, 412, 421
 Mobility management, 2, 8, 473–507 boundary crossings and, 486–490 cost analysis for (*see* Signaling cost analysis) performance evaluation, 490–504 performance evaluation framework, 484–486 schemes for, 479–484 special cases, 504–506
 Modified quadrature-overlapped convoluted cosine (MQOCC), 583, 586, 587f, 589t
 Modified quadrature-overlapped triangular cosine (MQOTC), 576–580, 586, 587f, 589t
 Moment generating function (MGF), 444, 454, 457, 464–465, 466–467
 Monte Carlo simulations, 626, 628f, 629f, 645
 Moore-Penrose pseudoinverse, 644
 Mote processor / radio (MPR) board, 25
 Mote sensor (MT) board, 25
 Movement-based mobility management, 473, 477–478 cost analysis for, 491–495 properties of, 480–481
MPLS. *See* Multi-protocol label switching
MPPE. *See* Microsoft point-to-point encryption
M-PSK. *See* *M*-ary phase shift keying
M-QAM. *See* *M*-ary quadrature amplitude modulation
MQOCC. *See* Modified quadrature-overlapped convoluted cosine
MQOTC. *See* Modified quadrature-overlapped triangular cosine
MRC. *See* Maximal ratio combining
MSC. *See* Mobile switching centers
M-sequences, 83, 91, 92, 94
MSK. *See* Minimum shift keying
MSKTC. *See* Minimum shift keying triangular cosine
M2TP. *See* Min-two Uniform Targets Protocol
MUD. *See* Multi-user detection
Multi-amplitude / phase modulation schemes, 562, 563, 568
 Multichannel Multipoint Distribution Service (MMDS), 332–333
 Multi-hop Cellular Network (MCN), 59, 60f, 61f
 Multi-hop communication systems, 443–470. *See also* Dual-hop systems; Non-regenerative communication systems; Regenerative communication systems relayed transmission classification, 445 relaying usefulness in, 445–448
 Multi-hop Radio Access Cellular (MRAC) scheme, 56
 Multi-hop relaying, 53–54, 55–60, 62, 66
 Multimedia Group Broadcast, 18
 Multimedia Messaging Service (MMS), 11, 15, 16, 17
 Multimedia Resource Function (MRF), 13
 Multipath channels, 120–127
 Multipath diversity, 139
 Multipath fading, 158, 160–161, 254
 Multipath interference (MI), 83, 86, 87, 89, 98–99, 109–111 isotropic free, 113, 114, 129–138, 139, 140–141 properties of, 100–101 techniques to combat, 102–109
 Multipath signal receiver, 124–125
 Multiple access interference (MAI), 83, 87, 89–90, 92, 93, 97, 98–99, 109–111 isotropic free, 113, 114–129, 131–133, 139, 140–141 MI compared with, 101 properties of, 99 techniques to combat, 102–109
 Multiple antennas, 613–649 channel capacity and (*see* Channel capacity) coding for, 634–639 suboptimum receiver interfaces and, 644–647

- Multiple-input, multiple output (MIMO)
system, 81, 106, 615, 616
channel capacity and (*see* Channel capacity)
DFE for, 669–671, 673–674
diagonal STBCs and, 657, 667–671,
673–674, 675, 677
- Multiple session handling, 8, 10
- Multi-protocol label switching (MPLS), 344
- Multi-service wireless Internet link
enhancements, 291–314
architecture of, 301–303
background and related work, 292–294
best-effort service interface, 306–308
performance with, 304–306
simulation setup, 295–297
- Multi-user detection (MUD), 83, 87,
103–104, 107, 108
- Nakagami fading
AOD and, 209, 211, 212, 218, 219–220,
228, 229, 230–233, 234f, 235f,
236
bivariate scenario, 230–232
dual-hop systems and, 450
- Nakagami / Nakagami fading, 213, 214–215
- Nakagami / Rician fading, 213, 214, 216
- NanoPeer Words, 37
- Nash equilibrium, 541, 551
- National (Inter-Radius) portability, 330–332
- Natural random mobility model, 392, 408
- Near-far effect, 100, 102–103, 157–158
- Network-based intrusion detection, 534
- Network resource attacks, 541
- Network Sensor Platform, 514
- Ninja project, 522
- No codes, 91
- Node activation, 406
- Node deactivation, 406, 407
- Node density, 394, 416
- Node mobility, 405, 407–409, 417–418, 421
- Node pairs, 399, 400
- Noise
additive white Gaussian (*see* Additive white Gaussian noise)
background, 430, 434
white, 158, 159
- Noise-enhancement problem, 104
- Noise-limited systems
CDMA evolution to, 132–133, 136–138
outage probability in, 450
uplink load and, 429
- Nokia, 56
- Non-cooperative games, 550–551
- Nonergodic Rayleigh fading channel,
626–631
- Non-IP access domain of Softswitch, 362
- Nonlinear interfaces, 645–647
- Non-overlapped pulse shaping, 592–593
- Non-regenerative communication systems,
444, 445, 447–448, 470
average BER in, 454–455, 456, 466–467
optimum power allocation in, 460–464,
468
outage probability in, 450–452, 464–466
- NTT DoCoMo, 560
- Nyquist receiver input filter (square-root),
658, 659
- Nyquist signalling rate, 614, 616
- ODMA. *See* Opportunity-Driven Multiple Access
- OFDM. *See* Orthogonal frequency division multiplexing
- Offset quadrature phase shift keying
(OQPSK), 561, 562, 564, 567, 569,
570, 586, 598, 599
- OMSK. *See* Overlapped minimum shift keying
- One-Pass With Advertising (OPWA),
313–314
- One-way hash functions, 340, 344
- OOBP. *See* Out-of-band power
- Open-loop power control, 83, 87, 102, 140,
159, 161
- Open-loop sector of up-link synchronization,
108, 109, 110f
- Open loop transmit diversity, 74
- Open Mobile Alliance (OMA), 12
- Opportunity-Driven Multiple Access
(ODMA), 57–58
- Optimum power allocation
in dual-hop systems, 458–464
in multi-hop systems, 468–470
- OPWA. *See* One-Pass With Advertising
- OQPSK. *See* Offset quadrature phase shift keying
- “Or of the downs” rule, 165, 166
- Orphan clusters, 406
- Orthogonal codes, 85–86, 91, 94–95, 113,
129, 138, 158, 642–643
- Orthogonal frequency division multiplexing
(OFDM), 633, 656
- Orthogonal variable spreading factor (OVSF)
codes, 83, 88, 91, 96, 113, 119
characteristics of, 93–94
up-link synchronization and, 109, 111
- Outage probability
in dual-hop systems, 450–454, 458, 460
in multi-hop systems, 444, 447, 464–466
multiple antennas and, 617, 626–627,
629, 631, 634
- Outer loop power control, 157, 158, 165, 200
- Out-of-band power (OOBP), 586–587
- Out-of-phase auto-correlation function
(ACF), 89, 90, 98, 99, 101, 113,
116
of complementary codes, 94, 96

- Out-of-phase auto-correlation function (ACF)
(continued)
 of Gold codes, 92
 of M-sequences, 91
 of OVSF codes, 94
 RAKE and, 105
 of Walsh-Hadamard sequences, 93
- Out-of-phase cross-correlation function (CCF), 93, 94, 96, 98, 112, 113
- Overlapped minimum shift keying (OMSK), 590, 597–599, 600f, 601f
- Overlapped pulse shaping, 592–593
- OVSF codes. *See* Orthogonal variable spreading factor codes
- Packet marking approach, 542
- Packet purse model, 539–540
- Packet trade model, 540
- Paging, 478, 484, 485, 486, 487, 493–494, 503
- Pairwise error probability (PEP), 635–636, 637, 638, 644, 645, 659, 664, 665, 677
- Parallel paging, 478
- Partial response waveforms, 562
- Password authentication protocol (PAP), 338
- Passwords, 533, 535, 536
- Payoff function, 551–552
- PCBC. *See* Power-control-based clustering
- PCS. *See* Personal Communications Service
- PDC. *See* Personal Digital Cellular technologies
- PDF. *See* Probability density function
- Peer-to-peer networking, 1–21. *See also* Mobile networks
- PEP. *See* Pairwise error probability
- Perfect power control, 160
- Periodic auto-correlation function (ACF), 89, 99, 101
- Periodic cross-correlation function (CCF), 89, 99, 112
- Personal Communications Service (PCS), 332–333
- Personal Computing and Communications (PCC) program, 54
- Personal Digital Cellular (PDC) technologies, 3
- Personalization of information and positioning, 2
- Pervasive computing, 513–528
 caching in (*see* Caching)
 challenges of, 515–517
 security in, 542–546
- PFR. *See* Probabilistic Forwarding Protocol
- PFS. *See* Pointer forwarding scheme
- PG values. *See* Processing gain values
- Phase-type distribution model, 504
- Piece-wise fitting methods, 563, 574–580, 605
- Pilot-aided Code Division Multiple Access (CDMA), 83, 107–108
- Pipe, 250–251, 258, 272
- Pointer forwarding scheme (PFS), 473, 477, 486. *See also* Two-level pointer forwarding scheme
 cost analysis for, 495–498
 properties of, 481
- Points in Calls (PICs), 366
- Point-to-point (PPP) protocol, 321, 325, 335, 337–338
- Point-to-Point Tunnelling Protocol (PPTP), 335, 337–339
- Poisson process, 478, 486, 487
- Pole capacity, 427, 428, 430, 432
- Pole equation, 428
- Policy Control Function (PCF), 13
- Portability architecture, 319–355
 elements of, 321–326
 key factors in, 326–327
 scenarios, 328–333
 system authentication, 320–321
 tunnelling for, 334–335
- Positioning, 2
- Post shaping processor, 564, 565, 566, 569, 572, 575
- Power allocation, optimum. *See* Optimum power allocation
- Power consumption
 clustering and, 416
 hybrid ad-hoc cellular networks and, 692
 multi-hop systems and, 446–447
- Power control, 179–203
 algorithm design, 201–203
 aspects of, 199–203
 closed-loop (*see* Closed-loop power control)
 command errors in, 162, 167–168
 distributed, 185
 disturbances, 185, 187–188
 downlink (*see* Downlink power control)
 error statistics, 190–191
 feasibility, 192–194
 feedback bandwidth, 186, 188–189
 feedback errors, 186
 fundamental algorithm, 184
 fundamental local limitations, 185–191
 global stability, 191–198, 203
 inner loop, 157, 158–159
 limited update rate, 186–187
 local stability, 186, 189–190
 measurement errors, 185, 187–188
 near-far effect vs., 102–103, 157–158
 open-loop, 83, 87, 102, 140, 159, 161
 outer loop, 157, 158, 165, 200
 perfect, 160
 rate reduction, 168–170, 174–175
 relation to literature, 184–185
 sensitivity function, 186

- Power control (*continued*)
step size, 163–164, 166
system model, 180–185
in third generation CDMA, 157–175
time delays, 186, 189–190
transmitter power limitations, 186, 190
update rate, 185
uplink, 165–166, 200–201
uplink load and feasibility of, 433–436,
 440
- Power-control-based clustering (PCBC),
 411–412, 421
- Power control command rate, 160–162
- Power control groups, 159, 162
- Power efficiency, 559, 561, 563, 567, 568
- Power gain, 180–182, 186–187, 192, 201
- Power spectral density (PSD), 566, 567, 570,
 571, 572, 574, 579, 600f, 601f
 expression derivation, 594–596
 numerical evaluation of, 596–597
 over linear channels, 597–598
 over non-linear channels, 598–599
 of single-sided bandwidth, 586
 time-domain convolution and, 580–585
- PPP. *See* Point-to-point protocol
- PPTP. *See* Point-to-Point Tunnelling Protocol
- Presence, 11, 17
- Presence Server (PC), 13
- Private keys, 533, 537–538, 544, 546
- Probabilistic Forwarding Protocol (PFR),
 41–47, 48
- Probability density function (PDF), 214, 230
- Processing gain (PG) values, 84, 87, 89, 95,
 96, 144–149
- Profiling, 514, 521
- Protocol Adapter, 363–364, 380
- Proxy-based adaptation, 521–522
- PSD. *See* Power spectral density
- PSTN. *See* Public Switched Telephone
 Networks
- Public key infrastructure (PKI), 543
- Public keys, 533–534, 537–539, 546, 549
- Public Switched Telephone Networks
 (PSTN), 474, 475f
- Pulse-shaped modems, 586–589
- Pulse shaping filters, 564, 565, 566, 571, 575
 piece-wise fitting and, 576–579
 spectral efficiency and, 592–593
- Pulse shaping waveforms (PSW), 559–606
 intuitive methods, 563, 569–574, 605
 piece-wise fitting methods, 563,
 574–580, 605
 roles of techniques, 564–569
 spectral efficiency of, 589–599
 time-domain convolution methods, 563,
 580–586, 591, 602, 605, 606
- 16QAM. *See* 16 Quadrature Amplitude
 Modulation
- QOCC. *See* Quadrature-overlapped
 convoluted cosine
- QO modulations. *See* Quadrature-overlapped
 modulations
- QORC. *See* Quadrature-overlapped raised
 cosine
- QOTC. *See* Quadrature-overlapped triangular
 cosine
- QPSK. *See* Quadrature phase shift keying
- 16 Quadrature Amplitude Modulation
 (16QAM), 76, 77, 78, 96, 676
- Quadrature-overlapped convoluted cosine
 (QOCC), 581–582, 584, 586–588,
 589t
- Quadrature-overlapped (QO) modulations,
 559–560, 562, 564, 581, 592–594,
 597–606
 additional examples of, 589–591
 features of importance, 573–574
 over linear channels, 597–598
 over non-linear channels, 598–599
- Quadrature-overlapped raised cosine
 (QORC), 567, 589t, 597–599,
 602–604
 intuitive methods and, 569, 571–572, 574
 MSK compared with, 573f
 piece-wise fitting and, 575–576
- Quadrature-overlapped triangular cosine
 (QOTC), 576–580, 582–583,
 586–587, 588f, 589t
- Quadrature phase shift keying (QPSK)
 diagonal STBCs and, 665, 676
 HS-DSCH and, 76, 77–78
 multiple antennas and, 638, 639f
 pulse shaping waveforms and, 561, 562,
 566, 567, 569, 570, 586, 592,
 599
 W-CDMA and, 5
- Qualcomm, 102
- Quality of Service (QoS)
 caching and, 526
 clustering and, 386
 for IMS, 9, 13, 14
 Internet link enhancements and, 291–292,
 308–309, 311–314
 for mobile networks, 6, 8
 mobility management and, 476, 485
 portability architecture and, 323, 326,
 328
- TCP and, 242
 for third generation CDMA, 158
 for third generation W-CDMA, 72, 75,
 77, 82
 uplink load and, 427, 428
- Quasi-orthogonal codes, 85–86, 91, 94, 138
- Queueing theory, 504
- RACH. *See* Random Access Channel
- Radio access, 4, 14

- Radio access controller, 322–324
 Radio access network (RAN), 5, 14. *See also*
 Universal Terrestrial Radio Access Network
 Radio Frequency ID (RFID), 514
 Radio link controller (RLC), 320
 Radio link protocol (RLP), 294, 320, 321,
 331
 Radio network controller (RNC), 320–321,
 324, 331, 332, 334–335, 427
 Radio resource management, 387–390
 RADIUS. *See* Remote Authentication Dial In User Service
 Raised-cosine triangular cosine (RCTC)
 modulation, 590, 597–599, 600f,
 601f, 602–604
 Raised-cosine waveform, 567, 569, 571–572,
 574, 575–576, 585, 591
 RAKE receivers, 83, 87, 90, 100, 120–121,
 125, 129–130, 132, 133, 137f
 equal gain combining, 105, 139–140
 maximal ratio combining, 105, 107, 108,
 130, 133, 139
 MUD and, 103
 properties of, 104–105
 RAN. *See* Radio access network
 Random Access Channel (RACH), 74
 Random error, 268, 270, 273, 274
 Random walk mobility model, 392
 Rank-and-determinant criterion, 636
 Rate gain, 615
 Rate-matching algorithms, 87, 88–89, 93,
 118–120
 Rational functions, 578
 Rayleigh fading
 AOD and, 211, 212, 218, 219, 220, 221,
 223, 224, 225, 227, 228, 229,
 232, 233f
 diagonal STBCs and, 658, 662
 dual-hop systems and, 450, 455
 ergodic, 619–626, 631
 frequency-flat, 616–617
 in independent and balanced channels,
 232
 multi-hop systems and, 443, 445, 465,
 470
 multiple antennas and, 616–617,
 619–631, 636
 nonergodic, 626–631
 TCP and, 254
 third generation CDMA and, 160, 173
 Rayleigh/Rayleigh fading, 214–215, 229
 RCTC modulation. *See* Raised-cosine triangular cosine modulation
 Reachability metric, 694
 Receiving walls, 36–40
 Recursive least-squares (RLS) algorithm, 656
 Reduced-state sequence estimation (RSSE),
 656, 657, 671, 675
 Reed-Muller code, 645, 646f
 Re-encryption server, 325
 Refresh time method, 518
 Regenerative communication systems, 444,
 445, 470
 average BER in, 455, 456, 467
 optimum power allocation in, 458–460,
 461–464, 468
 outage probability in, 451–452, 465–466
 Registration. *See* Location updates
 Registration area (RA), 474, 476, 479–480,
 481–482, 483–484, 495–496, 499
 Registration area (RA) boundary crossings,
 486–490, 501
 Registration area residence time (RRT), 486,
 487, 490–491, 495, 497, 498, 500,
 501, 502
 Release 4, 6
 Release 5, 6, 7, 8, 11, 71, 82
 evolution beyond, 79–81
 features of, 76–78
 Release 6, 6, 78, 79, 81
 Release 7, 81
 Release '99, 82
 features of, 71–75
 HSDPA compared with, 77–78
 Remote Antenna Units (RAU), 55, 56
 Remote Authentication Dial In User Service
 (RADIUS), 320–322, 324,
 330–332
 central server, 325–326
 local server, 325
 tunnelling and, 334–335
 Repetition codes (RCs), 662
 Replacement score, 519
 Repudiation of actions, 532
 Reputation, 551–552
 Residual life theorem, 488
 Residue Theorem, 490, 497, 502, 505
 Reverse half-cycle sinusoidal waveforms,
 585
 Reverse raised-cosine waveforms, 585
 RFID. *See* Radio Frequency ID
 Rich Voice, 15, 16
 Rician fading, 209, 211, 212, 218, 236
 Rician/Nakagami fading, 213, 214, 215
 RLC. *See* Radio link controllers
 RLP. *See* Radio link protocol
 RNC. *See* Radio network controllers
 Robust header compression (ROHC), 14
 Robustness
 of PFR, 44, 47
 against sampling timing jitter, 599–604
 Role based control, 543
 Rooftop, 56
 Rotated M -ary quadrature amplitude modulation (R-MQAM), 663–664
 Rotated MASK, constellations, 663–664
 Router-based intrusion detection, 535

- Routing
clustering and, 389–390
location-based, 690–695
mobile (see Mobile routers)
shortest-path, 389
source, 334
- RRT. *See* Registration area residence time
- RSSE. *See* Reduced-state sequence estimation
- SACCS. *See* Scalable Asynchronous Cache Consistency Scheme
- SACK Transmission Control Protocol (TCP), 241, 246, 249, 274
compared with other variants, 252–253
dynamics and performance in, 254–262, 264–269
maximum throughput, 269–270
properties of, 250–251
- Sampled Poisson process, 478
- Sampling timing jitter, 563, 567, 568, 587
QO modulation sensitivity to, 574
robustness against, 599–604
- SC. *See* Selection combining
- Scalability, 29
localized, 542–543
- Scalable Asynchronous Cache Consistency Scheme (SACCS), 527
- SCFQ. *See* Self Clocked Fair Queueing
- Scoreboard, 250, 252
- Secret keys, 533, 549
- Secured Hash Algorithm (SHA-1), 340, 344, 351–352, 354
- Secure efficient ad hoc distance vector (SEAD), 539
- Secure Network Encryption Protocol (SNEP), 548
- Security, 531–553
in ad hoc networks, 534, 536–542
distributed trust and, 543–544
enforcing cooperation, 539–541, 550–552
intrusion detection, 534–536
in mobile networks, 3, 8, 531–546
in pervasive computing, 542–546
protocol basics, 532–534
reputation and, 551–552
in sensor networks, 531–532, 546–552
Vigil system and, 544–546
- Selection combining (SC), 210–211, 453
CCI impact on, 218–234
comparison of algorithms, 219–227
fading impact on, 227–234
performance analysis of, 220–221
in RAKE, 105
- Selective paging, 478
- Selective Repeat, 296, 298, 299, 301, 305, 306
- Selective Slow Start (SSS), 276–277
- Self Clocked Fair Queueing (SCFQ), 302–303, 304
- Selfish and malicious nodes, 540
- Semi-blind relays, 452–453
- Sensor networks, 23–49
ad hoc networks compared with, 29
applications of, 28
data propagation protocols, 30–47
modeling aspects of, 24–28
new challenges for, 29–30
recent work in, 47–49
security in, 531–532, 546–552
- Sequence number attack, 536
- Sequential paging, 478
- Serial-to-parallel (S/P) converter, 615
- Served calls, 478
- Server resource attacks, 541
- Server-side adaptation, 521–522
- Service Agent, 364, 380
- Service Content Provision (SCP), 13
- Service Control Point (SCP), 366
- Service level agreement (SLA), 323, 326, 328, 333
- Service plane of Softswitch, 362
- Service Switching Function (SSF), 13
- Serving GPRS Support Node (SGSN), 242, 360
- Session hijacking, 536
- Session Initiation Protocol (SIP)
network architecture based on, 7–9, 12, 16, 17, 19
softswitch, 360, 362, 363, 369–372, 373f, 374–376, 377–379
- Session Initiation Protocol (SIP) Application Server (AS), 13
- SGSN. *See* Serving GPRS Support Node
- SHA-1. *See* HMAC-SHA-1; Secured Hash Algorithm
- Shadowing, 158, 161
- Shortest-path routing, 389
- Side lobe regeneration problem, 561, 563
- Siemens, 11
- Signal detection, 83, 107–108
- Signaling cost analysis, 478–479, 484–485, 490–504
for IS-41, 490–491
for movement-based mobility management, 491–495
for pointer forwarding, 495–498
for two-level pointer forwarding, 498–500
for two location algorithm, 500–504
- Signaling gateway, 360
- Signalling system 7 (SS7) standard, 476
- Signal-to-interference ratio (SIR), 180, 182–183, 184, 185, 188, 192–194, 197, 198, 202
- AOD and, 210, 211, 212, 216–217, 219, 233–234, 235–236

- Signal-to-interference ratio (SIR) (*continued*)
 best algorithm, 220, 223–224, 227, 236
 defined, 182
 downlink, 183, 200–201
 in third generation CDMA, 168, 175
 uplink, 183
- Signal-to-noise ratio (SNR)
 AOD and, 210
 diagonal STBCs and, 665, 676, 677
 dual-hop systems and, 449, 450,
 451–453, 454–455, 456, 457,
 459, 461, 463–464
 MUD and, 104
 multi-hop systems and, 447–448,
 464–465, 466, 468, 470
 multiple antennas and, 614, 618, 624,
 625, 627, 631, 633, 634, 636,
 637, 638
 pulse shaping waveforms and, 568
 RAKE and, 105
- Single-bit error, 188–190, 191, 197, 199
- Single-input-single-output (SISO) systems,
 671, 673–675
- Single-sided bandwidths, 586–587
- Sinks, 27–28, 31, 32, 33, 34, 37, 42, 47–48
- Sinusoidal quadrature overlapped triangular cosine (SQOTC), 591, 597–599,
 600f, 601f
- SIP. *See* Session Initiation Protocol
- SIR. *See* Signal-to-interference ratio
- SISO systems. *See* Single-input-single-output systems
- Site selection diversity transmit power control (SSDT), 171–172
- SLA. *See* Service level agreement
- Sleep-Awake Probabalistic Forwarding Protocol (SW-PFR), 48
- Sleeping nodes, 33–36, 39, 519, 541
- Small gain theorem, 196
- Small-set Kasami codes, 92
- Smart antennas, 78, 80, 106, 140–141
- Smart dust clouds, 27, 36–39
- Smart dust particles, 26–28, 48
- Smart Dust Project, 24
- Smart sensors, 514
- Smart spaces, 516–517, 542, 544, 545
- Smith predictor, 190, 201
- SNDCP. *See* Subnetwork Dependent Convergence Protocol
- SNEP. *See* Secure Network Encryption Protocol
- Snoop protocol, 248, 258, 283, 294, 296. *See also* Berkeley Snoop
- SNR. *See* Signal-to-noise ratio
- Soft capacity, 429–430
- Soft handoff, 157, 158, 172
 downlink power control during, 166–171
 uplink power control during, 165–166
- Soft handover, 79–80, 199–200, 437, 439
- Soft resources, 426
- Softswitch, 359–380
 call model, 365–366
 CCL architecture for, 363–364, 365f, 380
 ISC architecture for, 361–362, 379
 message flows of call control, 369–379
- Softswitch Service Provider Interface (SPI),
 363, 364, 367–368, 369–376, 380
- Source routing, 334
- Space-time block codes (STBCs), 634. *See also* Diagonal space-time block codes
- Space-time codes, 106
 CSI unavailability and, 643–644
 decoding, 639
 layered, 634
 multiple antennas and, 616, 634,
 637–638, 639–644
 schemes, 639–644
 trellis, 634, 643, 656
- Sparse Topology and Energy Management (STEM) Protocol, 33–36
- Spatial diversity, 57–58
- Spatial interference, 614
- Spatial multiplexing, 615
- SPI. *See* Softswitch Service Provider Interface
- Spoofing, 536, 541
- Spreading codes, 87, 94, 96, 98, 107, 109,
 119, 141
 isotropic, 113, 129, 130
- Spreading efficiency (SE), 127
- Spreading factors (SF), 93
- SQOTC. *See* Sinusoidal quadrature overlapped triangular cosine
- Square waveforms, 585
- SSDT. *See* Site selection diversity transmit power control
- Standard interference function, 194–195
- Standard iterative algorithm, 195
- Stateful approach to cache consistency, 527
- Stateless approach to cache consistency, 527
- STEM Protocol. *See* Sparse Topology and Energy Management Protocol
- Strategy set, 552
- Subnetwork Dependent Convergence Protocol (SNDCP), 242
- Sum of Hyper-Exponential (SOHYP) model, 504
- Surrogate machines, 514
- SW-PFR. *See* Sleep-Awake Probabalistic Forwarding Protocol
- Symmetric cryptography / encryption, 340,
 344, 533, 539
- TCP. *See* Transmission Control Protocol
- TDC. *See* Time Delay Compensation; Time-domain convolution methods
- TDD. *See* Time Division Duplex

- TD-LAS, 96–97
TDMA. *See* Time Division Multiple Access
TD-SCDMA. *See* Time Division Synchronous Code Division Multiple Access
Telecom Management, 13
Temporary-location directory number (TLDN), 476
Terminal identification, 8
Terminal Management, 13
 μ TESLA protocol, 548–549
TH-CDMA. *See* Time hopping Code Division Multiple Access
Third generation Code Division Multiple Access (CDMA), 157–175. *See also* Power control
compressed mode, 173–174
reduced active set, 172–173
Third generation mobile all-IP network Softswitch. *See* Softswitch
Third Generation Partnership Project (3GPP), 8, 12, 14, 57, 71
Third generation Wideband Code Division Multiple Access (W-CDMA), 71–82
commercial deployment, 71–75
performance enhancement, 78
Third-party call control, 368
Three-phase multiple access (TPMA), 409
Threshold cryptography, 537–538
Throughput fairness, 254–255, 260, 261f, 267, 274
TIA/EIA, 84
Time-based location update, 477–478, 480
Time Delay Compensation (TDC), 190, 196, 197, 198, 201
Time Division Duplex (TDD), 53, 56, 57, 62, 66, 161
Time Division Multiple Access (TDMA), 4, 86, 104
Time Division Synchronous Code Division Multiple Access (TD-SCDMA), 84, 93, 103, 108–109
Time-domain convolution (TDC) methods, 563, 580–586, 591, 602, 605, 606
Time hopping Code Division Multiple Access (TH-CDMA), 84, 85, 87
Timeout events, 246
Timer expiration, 406, 407
Time-shift operator, 184
TLA. *See* Two location algorithm
TLDN. *See* Temporary-location directory number
TLPDS. *See* Two-level pointer forwarding scheme
TPC command. *See* Transmit power control command
Transmission Control Protocol (TCP), 241–285, 293–294, 295–299, 301, 307–308, 313
basic variants, 249–252
comparison among variants, 252–254
dynamics and performance in, 254–269
FACK (*see* FACK Transmission Control Protocol)
ideal end-to-end, 269–270
indirect, 280, 283
maximum throughput, 269–270
modification of variants, 270–283
multi-service link layers and, 304–306
receiver-only modifications, 277–278
SACK (*see* SACK Transmission Control Protocol)
sender and receiver modifications, 278–280
sender-only modifications, 270–277
session hijacking in, 536
single-service link layers and, 297–299
wireless, 278–279
Transmission Control Protocol / Internet Protocol (TCP/IP), 241, 242
Transmission Control Protocol (TCP)
New-Reno, 241, 248, 251
compared with other variants, 252–253
dynamics and performance in, 254–269
maximum throughput, 269–270
properties of, 250
Transmission Control Protocol (TCP)
OldTahoe, 248
Transmission Control Protocol (TCP) Reno, 241, 248, 274
dynamics and performance in, 254–262, 264–269
properties of, 249–250
Transmission Control Protocol (TCP) Santa Cruz, 274–276
Transmission Control Protocol (TCP) Tahoe, 241, 248
compared with other variants, 252
dynamics and performance in, 254–262, 264–269
properties of, 249
Transmission Control Protocol (TCP) Vegas
compared with other variants, 253
dynamics and performance in, 254–262, 264–269
properties of, 251–252
Transmission Control Protocol Westwood (TCPW), 273–274
Transmission power, 394
Transmit power control (TPC) command, 162, 173
base station synchronization with, 168
error rate, 168–170, 171
rate reduction, 168–170
site selection diversity, 171–172

- Transmit power control (TPC) command
(continued)
 soft handoff and, 165–166
 step size, 163–164
- Transport level approaches
 comparison of, 284t
 non-end-to-end, 280–283
- Transport plane of softswitch, 361–362
- Transport Signalling Gateway (TSGW), 13
- Trellis space-time codes, 634, 643, 656
- Triangular waveforms, 585
- Trigonometric functions, 578
- Triple Data Encryption Standard (3DES),
 346–348, 353–355
- Triple-hop systems, 464, 465
- Tunnelling, 333–344
 compulsory, 336–337
 layer 2 (*see* Layer two tunnelling)
 layer 3, 340–344
 MPLS, 344
 source routing *vs.*, 334
 voluntary, 336, 337f
- Tunnelling client, 324–325
- Tunnelling server, 325
- Two-hop clusters, 386, 401, 411
- Two-level location management scheme, 473
- Two-level pointer forwarding scheme
 (TLPDS)
 cost analysis for, 498–500
 properties of, 481–483
- Two location algorithm (TLA), 473, 477
 cost analysis for, 500–504
 properties of, 483–484
- Ubiquitous computing (UC), 1, 2, 19–20
- Ubiquitous national portability, 332–333
- UDP. *See* User Datagram protocol
- Ultralow Power Wireless Sensor Project, 25
- Ultra-wideband (UWB) technology, 84–85,
 563, 605
- Umbrella routers, 616
- UMTS. *See* Universal Mobile
 Telecommunication System
- Unauthorized use of resources, 532
- Uncovered nodes, 402
- Uniform wireless channel error, 255–262,
 267–268
- Unitary codes, 94, 96
- Unitary space-time modulation, 643–644
- Unity load, 196
- Universal Mobile Telecommunication
 System (UMTS), 1, 4, 6–7, 20, 75,
 77, 359–380. *See also* Softswitch
 A-Cell relay and, 53, 57–58, 66
 extending standardization scope of, 12
 portability architecture and, 324
 power control and, 160, 162
- Universal Mobile Telecommunication
 System / Frequency Division
 Duplex (UMTS / FDD), 4–5
- Universal Mobile Telecommunication
 System Radio Access (UTRA), 4,
 84, 88, 93, 103, 108, 119
- Universal Mobile Telecommunication
 System / Time Division Duplex
 (UMTS / TDD), 5
- Universal Terrestrial Radio Access Network
 (UTRAN), 5, 14, 80, 82, 360
- Universal Wireless Communications
 (UWC-136), 3, 4
- University of California at Berkeley, 24, 25,
 514, 522, 548
- University of California at Los Angeles
 (UCLA), 25
- Uplink fractional load, 427
- Uplink load, 425–440
 capacity and, 427–431
 centralized estimates, 427, 433–439, 440
 decentralized estimates, 426, 431–433,
 439–440
- Uplink noise rise, 427
- Uplink power control, 165–166, 200–201
- Uplink relative load, 425, 426, 427, 430, 432,
 437, 438, 439
 defined, 429
 power control feasibility and, 433–436,
 440
- Uplink signal-to-interference ratio (SIR), 183
- Up-link synchronization control, 83, 87,
 108–109, 110f, 111
- Useful degree of a neighbor, 692–693
- User authentication, 8
- User Datagram protocol (UDP), 292–294,
 295–296, 300–301, 304–306,
 307–308, 336
- User mobility, 246
- UTRA. *See* Universal Mobile
 Telecommunication System Radio
 Access
- UTRAN. *See* Universal Terrestrial Radio
 Access Network
- UWB. *See* Ultra-wideband technology
- UWC-136. *See* Universal Wireless
 Communications
- Valid scope area, 520
- Variable gain relays, 452–453
- Variable Transmission Range Protocol
 (VTRP), 48
- Video services, 6–7, 9
- Vigil system, 544–546
- Virtual links, 302
- Virtual private network (VPN) tunnels, 319,
 338, 344

- Visitor Location Register (VLR), 474, 476, 477, 478, 480, 481–482, 491, 495, 499, 503
- Viterbi algorithm, 671
- VLR. *See* Visitor Location Register
- Voice over IP (VoIP), 19, 74, 360, 362, 364, 367, 370, 371, 379
- Voice services, 6–7, 8, 9, 15, 16, 172, 173
- Voluntary tunnels, 336, 337f
- VPN tunnels. *See* Virtual private network tunnels
- VTRP. *See* Variable Transmission Range Protocol
- Walsh-Hadamard matrix, 93, 94
- Walsh-Hadamard sequences, 83, 86, 91, 94, 96, 113, 129
- CC-CDMA and, 116–117, 119
 - characteristics of, 93
 - DS-CDMA and, 134, 136, 138
- Water-filling equation, 618, 629
- Waveform shaping techniques. *See* Pulse shaping waveforms
- W-CDMA. *See* Wideband Code Division Multiple Access
- WDS. *See* Wired Server
- “What I want” type of information, 513, 515, 523
- White noise, 158, 159. *See also* Additive white Gaussian noise
- Wide-area cellular wireless networks, 241–285. *See also* Transmission Control Protocol
- Wideband Code Division Multiple Access (W-CDMA), 84, 88, 119
- A-Cell relay and, 53, 58, 62, 64–65, 66
 - described, 4–5
 - Gold codes and, 92
 - isotropic spreading techniques and, 111
 - MUD and, 103
 - OVSF codes and, 93
 - power control in, 186, 187–188, 189, 190, 200
 - third generation evolution, 71–82
 - uplink load in (*see* Uplink load)
 - up-link synchronization and, 108
- WiFi networks, 58–59
- Window scheme, 519
- Wired-cum-wireless networks, 269, 278
- Wired Server (WDS), 295, 296, 297, 299, 300, 304
- Wireless Client (WLC), 295, 296, 297, 299, 300, 304
- Wireless Integrated Network Sensors (WINS) Project, 25
- Wireless link bandwidth, 255–267
- Wireless Local Area Networks (WLAN), 5–6, 20, 295
- Wireless Personal Area Networks (WPAN), 20
- Wireless sensor networks. *See* Sensor networks
- Wireless Transmission Control Protocol (WTCP), 278–279, 282–283
- WLAN. *See* Wireless Local Area Networks
- WLC. *See* Wireless Client
- WPAN. *See* Wireless Personal Area Networks
- WTCP. *See* Wireless Transmission Control Protocol
- W3C, 12
- Zero forcing (ZF)-BLAST interface, 645, 646–647
- Zero forcing (ZF) interface, 644
- Zero load, 196
- Zero-outage capacity, 617