FUNDAMENTALS OF MOBILE DATA NETWORKS

GUOWANG MIAO, JENS ZANDER,
KI WON SUNG AND SLIMANE BEN SLIMANE

Fundamentals of Mobile Data Networks

This unique text provides a comprehensive and systematic introduction to the theory and practice of mobile data networks.

Covering basic design principles as well as analytical tools for network performance evaluation, and with a focus on system-level resource management, you will learn how state-of-the-art network design can enable you flexibly and efficiently to manage and trade off various resources such as spectrum, energy, and infrastructure investments.

Topics covered range from traditional elements such as medium access, cell deployment, capacity, handover, and interference management, to more recent cutting-edge topics such as heterogeneous networks, energy- and cost-efficient network design, and a detailed introduction to 4G Long Term Evolution (LTE).

Numerous worked examples and exercises illustrate the key theoretical concepts and help you put your knowledge into practice, making this an essential resource whether you are a student, researcher, or practicing engineer.

Guowang Miao is an Associate Professor at KTH Royal Institute of Technology. After receiving his PhD from Georgia Institute of Technology, he spent two years working in industry as a Senior Standard Engineer and 3GPP delegate at Samsung Telecom America in Dallas and was awarded an Individual Gold Award.

Jens Zander is a full Professor, and co-founder and Scientific Director of Wireless@KTH, at KTH Royal Institute of Technology. He is on the board of directors of the Swedish National Post and Telecom Agency (PTS) and a member of the Royal Academy of Engineering Sciences.

Ki Won Sung is a Docent in the Communications Systems Department at KTH Royal Institute of Technology. He is also affiliated with KTH Center for Wireless Systems (Wireless@KTH).

Slimane Ben Slimane is an Associate Professor in the Communication Systems Department at KTH Royal Institute of Technology, having previously been an Assistant Professor in the Department of Signals, Sensors, and Systems.

Fundamentals of Mobile Data Networks

GUOWANG MIAO
KTH Royal Institute of Technology

JENS ZANDER
KTH Royal Institute of Technology

KI WON SUNG
KTH Royal Institute of Technology

SLIMANE BEN SLIMANE
KTH Royal Institute of Technology

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781107143210

First published 2016

Printed in the United Kingdom by TJ International Ltd. Padstow, Cornwall

A catalogue record for this publication is available from the British Library

Library of Congress Cataloguing in Publication data
Names: Miao, Guowang, author. | Zander, Jens, author. | Sung, Ki Won,
author. | Slimane, Slimane Ben, author.
Title: Fundamentals of mobile data networks / Guowang Miao, Jens Zander,
Ki Won Sung, and Slimane Ben Slimane.
Description: Cambridge University Press, 2016.
Identifiers: LCCN 2015041666 | ISBN 9781107143210 (hardback)
Subjects: LCSH: Wireless communication systems. |
Cell phone systems. | Data transmission systems.
Classification: LCC TK5103.2.M49 2016 | DDC 621.3845/6–dc23
LC record available at http://lccn.loc.gov/2015041666

ISBN 978-1-107-14321-0 Hardback

Additional resources for this publication at www.cambridge.org/9781107143210

Contents

Preface *page* x
Acronyms xii
Notations xv

1 Introduction 1
1.1 Historical perspective on radio resource management 1
1.2 Key problems in wireless systems 2
1.2.1 Path loss—the early days 3
1.2.2 Thermal noise 4
1.2.3 Interference—the limited spectrum 4
1.2.4 Infrastructure cost and energy consumption 6
1.3 Wireless access networks—the issues 6
1.4 Outline of the book 10
References 11

2 Wireless network models 12
2.1 Introduction 12
2.2 Models for wireless access networks 14
2.3 Service scenarios and performance metrics 16
2.4 Radio resource management in wireless access networks 18
2.4.1 Orthogonal signal sets 22
2.4.2 Guaranteed service quality—blocking 22
2.4.3 Best effort—non-blocking 23
Exercises 25
References 26

3 Medium access control 27
3.1 Overview 27
3.2 Data traffic and performance measures 28
3.2.1 Delay 29
3.2.2 Delivery performance 30
3.2.3 Throughput 31
3.3 Contention-free access protocols 32
3.3.1 Resource assignment techniques 32

3.3.2 Dynamic access protocols 41
3.4 Contention-based access protocols 44
3.4.1 ALOHA 44
3.4.2 Carrier sense multiple access 50
3.4.3 CSMA with collision detection 52
3.4.4 Carrier sense multiple access with collision avoidance 52
3.5 Applications 60
3.5.1 IEEE 802.11 60
3.5.2 Cellular networks 62
Exercises 63
References 64

4 Scheduling 65
4.1 Introduction 65
4.2 Issues in wireless scheduling 68
4.2.1 Quality of service 68
4.2.2 Channel variation 69
4.3 Wireless scheduling and capacity region 70
4.3.1 Uplink multi-user capacity 72
4.3.2 Downlink multi-user capacity 74
4.4 Round-robin scheduling 76
4.5 Max throughput scheduling 76
4.6 Proportional fair scheduling 78
4.7 Max–min scheduling 80
4.8 Max utility scheduling 81
4.9 Scheduling in OFDMA systems 87
4.9.1 Max throughput scheduling in OFDMA 88
Exercises 93
References 94

5 Principles of cellular systems 95
5.1 Introduction 95
5.2 Orthogonal multiple access cellular systems 95
5.2.1 Coverage planning 95
5.2.2 Static channel allocation 97
5.2.3 Capacity analysis—guaranteed service 102
5.2.4 Traffic-based capacity analysis 106
5.2.5 Best-effort data services 110
5.2.6 Outage-based capacity analysis 114
5.2.7 Directional antennas and sectorizations 120
5.3 CDMA cellular systems 122
5.3.1 Uplink capacity of DS-CDMA systems 125
5.3.2 Traffic-based capacity of DS-CDMA systems 131
5.3.3 Downlink capacity of DS-CDMA systems 133

5.3.4 Multi-service DS-CDMA systems 134
Exercises 136
References 144

6 **Transmitter power control** 146
6.1 Introduction 146
6.2 Performance metric and conditions of achievability 146
6.3 Centralized power control 149
6.3.1 SIR balancing 150
6.3.2 Admission control 152
6.4 Distributed power control 153
6.4.1 Iterative power control 153
6.4.2 Convergence 154
6.4.3 General sufficient conditions for convergence 154
6.4.4 Distributed power control with power constraints 157
6.4.5 Admission control 158
6.4.6 Dynamics of power control 158
6.5 Power control for elastic traffic 159
6.5.1 Achievable region 159
6.5.2 Distributed power control for wireless data 162
6.6 Power control in DS-CDMA cellular systems 166
Exercises 171
References 175

7 **Interference management** 177
7.1 Classification of interference management techniques 177
7.1.1 Interference management categories 178
7.1.2 Key elements of interference management 179
7.2 Interference avoidance 181
7.2.1 Reuse partitioning 181
7.2.2 Multi-cell scheduling 184
7.3 Interference randomization 186
7.4 Interference cancellation 189
7.4.1 Successive interference cancellation 189
7.4.2 Transmit beamforming 190
7.5 Interference management for heterogeneous networks 192
Exercises 194
References 197

8 **Association and handover** 199
8.1 Anatomy of handover 199
8.1.1 Location management and handover 199
8.1.2 Types of handover 200
8.1.3 Handover phases 201

8.2 The handover decision problem 202
8.2.1 Performance metric 202
8.2.2 Tradeoff between handover frequency and failure 203
8.2.3 Impact of handover criteria 206
8.2.4 An example handover decision algorithm 208
8.3 Handover resource management 209
8.4 Soft handover 213
8.4.1 Soft handover procedure in practical systems 214
8.4.2 Fade margin improvement 216
8.4.3 Effects of soft handover on DS-CDMA capacity 218
8.5 User association 221
8.5.1 Load balancing 221
8.5.2 Association in heterogeneous networks 223
Exercises 224
References 226

9 Energy-efficient design 227
9.1 Introduction 227
9.2 Energy consumption in wireless networks 228
9.3 Energy-efficient transmission 229
9.3.1 Ideal transmission 230
9.3.2 Energy-efficient transmission in practice 231
9.3.3 Energy-efficient transmission in frequency-selective channels 235
9.4 Tradeoff in network resource utilization 241
9.4.1 Energy and spectral efficiency in interference-free channels 241
9.4.2 Energy and spectral efficiency in interference channels 241
9.5 Energy-efficient MAC design 245
9.6 Energy-efficient network management 250
9.6.1 Energy-efficient network deployment 250
9.6.2 Heterogeneous network deployment 254
9.6.3 Energy-efficient cellular network operation 255
Exercises 256
References 257

10 Long term evolution 258
10.1 Physical layer for downlink 258
10.1.1 Orthogonal frequency division multiplexing 258
10.1.2 Orthogonal frequency division multiple access 261
10.1.3 Multiple antenna techniques 262
10.2 Physical layer for uplink 267
10.2.1 Basics of SC-FDMA 267
10.2.2 SC-FDMA parameters for LTE 269
10.2.3 LTE random access 270
10.3 Interference management in LTE 272

10.3.1 Soft frequency reuse 272
10.3.2 Coordinated multi-point transmission 277
Exercises 279
References 281

11 Wireless infrastructure economics 282
11.1 Communication infrastructures 282
11.2 Wireless access economics 286
11.3 Spectrum cost and regulation 290
11.4 Affordable wideband wireless access 291
Exercises 297
References 297

About the authors 299

Index 302

Preface

The world has seen astonishing developments in wireless communications. From the early days when wireless was seen as a new and complex technology that required skilled operators to work, to a situation where wireless has become a truly pervasive technology with devices in everyone's pocket. Voice communications, including mobile telephony, have dominated the first century of wireless communications. The technical challenges have been dominated by the struggle of the engineer against nature—how to facilitate communications over long distances and how to overcome adverse radio propagation conditions. With the advent of digital communications, we have over recent decades seen marvelous advances in this area, with technologies such as error control coding, digital signal processing, advanced antenna technologies and others. Meanwhile, the number of wireless users has skyrocketed. In addition we now witness wireless Internet access becoming a dominant technology for all kinds of IT services. A necessary prerequisite for this development is that wireless access is abundant and becomes almost free. The consequence is that data rates in wireless communications have increased dramatically during the last decade. The industry predicts an exponential increase of data traffic that would correspond to a 1000-fold increase in traffic between 2010 and 2020. It has become obvious that traditional measures for increasing data rates in the wireless links, e.g. coding and signal processing, are not going to save the day since these techniques now operate close to their theoretical limits, regardless of their complexity. Instead, much of the focus of the engineering work has shifted to what can be seen as the social struggle for scarce resources. The proper management of resources such as frequency spectrum, energy consumption, and to a large extent monetary investments in infrastructure (base stations and the like) is now a key issue. The design objectives have changed from "how can we provide high quality communications in a single radio link?" to "how can we create sustainable systems that provide affordable high quality wireless communications for billions of users?" The latter question is mainly one of Radio Resource Management (RRM), which is the main theme of this book. The book approaches this problem in the following way. First we study various aspect of the classic RRM problem in wireless networks: given certain resources, a certain infrastructure of access point/base stations and a frequency spectrum, how can we maximize the capacity of the system, e.g. in terms of the number of users or the data rate per unit area. The key issue that is handled in this part is the complex mutual interference between the various network elements. In the second part, we address the problem of how, and how much, infrastructure should be deployed to meet certain user demand. As we will see there are no theoretical limits to the wireless data rate that can be provided, the problem becomes more about cost. The question becomes: “How can we meet customer expectations at the lowest cost?” where the cost for infrastructure, spectrum, energy, and so on are taken into account. Throughout the book, examples from state-of-the-art technologies such as LTE and recent WiFi standards are provided.

The book is intended as a textbook for a graduate course in wireless networks. The reader should be familiar with the fundamentals of radio communications and digital communications. Some basic queuing theory can also be useful. Wireless networks are complicated systems, which makes the design and performance analysis inherently difficult. Several approaches are taken in the book. Classical analysis involves highly simplified models but renders easily tractable results. Slightly more elaborate models are analyzed by means of numerical analysis. The book contains plenty of worked examples, figures and homework exercises in each chapter. Some of the examples and exercises require simple simulations. In addition, the slides and exercise solutions are also available for course teachers. At the end of each chapter, a small list of references is provided. We have no intention to exhaust the immense research literature and only those that are very closely related to the book material are given. All the material has been used in the courses given by the authors at KTH Royal Institute of Technology in Stockholm, Sweden.

The authors would like to thank the selfless help received in the development of the early versions of the manuscript. We are grateful for the contributions of Claes Beckman, Goran Anderson, Amin Azari, Peiliang Chang and Yanpeng Yang. In particular, we want to thank the several generations of graduate students at KTH, who have been instrumental in solving many of the problems and proofreading the manuscript. Their suggestions were extremely valuable in correcting typos and identifying weaknesses. Last but not least, we also appreciate the anonymous reviewers for providing valuable and in-depth comments on the early draft, which helped us improve the book coverage and greatly strengthened the final book.

Stockholm
Guowang Miao
Ki Won Sung
Slimane Ben Slimane
Jens Zander

Acronyms

3GPP	3rd Generation Partnership Project
ACK	ACKnowledgment
AP	Access Point
ARQ	Automatic Repeat reQuest
AWGN	Additive White Gaussian Noise
BER	Bit Error Rate
BLER	BLock Error Rate
bps	bits per second
BPSK	Binary Phase Shift Keying
BS	Base Station
C/I	Carrier to Interference
CDF	Cumulative Distribution Function
CDMA	Code Division Multiple Access
CoMP	Coordinated Multi-Point Transmission
CP	Cyclic Prefix
CRA	Conflict Resolution Algorithm
CRE	Cell Range Extension
CRPC	Constant Received Power Control
CSI	Channel State Information
CSG	Closed Subscriber Group
CSMA	Carrier Sense Multiple Access
CSMA/CA	Carrier Sense Multiple Access with Collision Avoidance
CSMA/CD	Carrier Sense Multiple Access with Collision Detection
CTS	Clear to Send
CW	Contention Window
DCF	Distributed Coordinated Function
DCH	Dedicated CHannel
DCPC	Distributed Constrained Power Control
DIFS	DCF Inter-frame Space
DPC	Distributed Power Control
DS-CDMA	Direct Sequence-Code Division Multiple Access
DSP	Digital Signal Processing
DSSS	Direct Sequence Spread Spectrum
EE	Energy Efficiency

FDD	Frequency Division Duplex
FDMA	Frequency Division Multiple Access
FFT	Fast Fourier Transform
FH-CDMA	Frequency Hop-Code Division Multiple Access
FSTD	Frequency Switched Transmit Diversity
GoS	Grade of Service
GPRS	General Packet Radio Service
GSM	Global System for Mobile communications
HARQ	Hybrid Automatic Repeat reQuest
HetNet	Heterogeneous Networks
ICIC	Inter-Cell Interference Coordination
ICT	Information and Communication Technology
IFFT	Inverse Fast Fourier Transform
IP	Internet Protocol
ITU	International Telecommunication Union
LTE	Long-Term Evolution
MAC	Medium Access Control
MAHO	Mobile-Assisted Handover
MCS	Modulation and Coding Scheme
MIMO	Multiple Input Multiple Output
ML	Maximum Likelihood
M-QAM	M-ary Quadrature Amplitude Modulation
MRC	Maximum Ratio Combining
MU-MIMO	Multiple-User MIMO
NACK	Negative ACKnowledgment
NAV	Network Allocation Vector
OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal Frequency Division Multiple Access
PAPR	Peak to Average Power Ratio
PDCCH	Physical Downlink Control CHannel
PDSCH	Physical Downlink Shared CHannel
PER	Packet Error Rate
PF	Proportional Fair
PG	Processing Gain
PN	Pseudonoise
PRB	Physical Resource Block
P/S	Parallel to Serial
QAM	Quadrature Amplitude Modulation
QoS	Quality of Service
QPSK	Quadrature Phase Shift Keying
RACH	Random Access CHannel
RAP	Radio Access Point
RAT	Radio Access Technology
RB	Resource Block

RE	Resource Element
RF	Radio Frequency
RR	Round-Robin
RRM	Radio Resource Management
RSRP	Reference Signal Received Power
RSRQ	Reference Signal Received Quality
RSS	Received Signal Strength
RSSI	Received Signal Strength Indicator
RTS	Request To Send
SC-FDMA	Single Carrier-Frequency Division Multiple Access
SDMA	Space Division Multiple Access
SE	Spectral Efficiency
SFBC	Space Frequecy Block Code
SIC	Successive Interference Cancelation
SINR	Signal to Interference Plus Noise Ratio
SIR	Signal to Interference Ratio
S-MAC	Sensor-Medium Access Control
SNR	Signal to Noise Ratio
S/P	Serial to Parallel
SU-MIMO	Single-User MIMO
SVD	Singular Value Decomposition
TDD	Time Division Duplex
TDMA	Time Division Multiple Access
TPC	Transmitter Power Control
TTI	Transmission Time Interval
UE	User Equipment
UMTS	Universal Mobile Telecommunications System
VoIP	Voice over IP
WLAN	Wireless Local Area Network

Notations

Scalars

A	Area
C	Channel capacity
D	Delay
D_o	Cell radius
d	Distance
E	Energy
g	Link power gain
G	Number of guard samples
h	Normalized link power gain
H	Channel signal gain
I	Scheduling indicator
N_{BS}	BS density
N_0	Noise power spectral density
P	Radio power at either the sender or receiver side
P_c	Circuit power
P_l	Listening circuit power
P_s	Sleep circuit power
$\hat{P}$	Maximum power
p	Probability
Q	Number of bits
q	Probability
R	Data rate
r	Data rate
S	Throughput
T	Duration
t	Time
U	Utility
u_e	Energy efficiency
u_s	Spectral efficiency
w	Size of a contention window
W	Signal bandwidth
η	Normalized thermal noise

Γ	SINR
κ	Cross-correlations between two waveforms
θ	Gap between the channel capacity and a practical coding and modulation scheme
γ	SINR target
δ	Access attempt rate
χ	Ratio between the channel gain and the interference plus noise
γ_0	Target threshold
ν	Assignment failure rate
ω	Active terminals per area unit
λ	Arrival or dropping rate
σ	Rate of retransmission
ξ	Expectation of retransmission delay
ζ	Power amplifier efficiency
ϕ	Network coupling factor
α	Path loss exponent

Vectors and matrices

$\mathbf{G}$	Link power gain matrix
$\mathbf{H}$	Normalized link power gain matrix
$\mathbf{h}$	Channel signal gain matrix
$\mathbf{I}$	Identity matrix
$\mathbf{N}$	Noise vector
$\mathbf{P}$	Power vector
$\mathbf{R}$	Data rate vector

Others

$\mathcal{C}$	Capacity region
$E[X]$	Expectation of X
$f()$	Mapping from SINR to data rate for a link
$\mathbb{I}$	Interference function
$i[t]$	The index of the user scheduled at time t
$\rho(X)$	Dominant eigenvalue of X
X_i	X of user i
X_{ij}	X between i and j
$X[t]$	X at time t
X^*	Optimal value of X
X'	First-order derivative of X
$X(Z)$	One-to-one mapping from Z to X

$X(\mathbf{Z})$	N-to-one mapping from $\mathbf{Z}$ to X
$\mathbf{X}(\mathbf{Z})$	N-to-M mapping from $\mathbf{Z}$ to $\mathbf{X}$
$\mathbf{X}^T$	Transpose of $\mathbf{X}$
$\mathbf{X}^H$	Hermitian of $\mathbf{X}$

1 Introduction

1.1 Historical perspective on radio resource management

As J. C. Maxwell had predicted in the 1850s, wireless transmission of electrical energy was feasible. Several decades later Heinrich Hertz managed to experimentally verify Maxwell's daring ideas with his award-winning experiment in 1888. He was able to demonstrate that his 600 MHz transmitter was capable of producing a spark in his simple receiver a few meters away in his laboratory. Although several scientists and inventors would like to claim the fame of inventing radio as we know it today, it took an engineer to bring this groundbreaking research into practical use. The Italian pioneer Guglielmo Marconi was the first to make practical and commercial use of the so-called Hertzian waves. After some initial experiments on his father's estate in 1895, his wireless apparatus gradually became a commercial success. It eventually made Marconi the first, but certainly not the last, millionaire in the wireless business. From humble beginnings, transmitting messages a few hundred meters in his first experiments, in 1901 he was finally able to demonstrate wireless communication across the Atlantic Ocean from Poldhu in Cornwall, England, to Newfoundland, Canada. In the decades to follow, wireless communications became an essential technology onboard ships. The early 1920s saw the advent of radio broadcasting, bringing wireless receivers into every home. We know what happened later—wireless has created a deep impact in our daily lives through success stories such as TV broadcasting, worldwide shortwave communication, satellite communications, and in recent decades mobile telephony and wireless and mobile Internet access.

The latest chapter in this story started to be written in the early 1980s with the commercial success of automated mobile telephony and mobile data. Examples of so-called first generation mobile telephone systems are the NMT system in Scandinavia (1981), AMPS in the USA (1984), TACS in the UK (1984) and other systems. These systems were targeting limited markets, terminals were expensive and they never reached very high user penetrations. The first-generation systems were analog designs—only the switching logic relied on digital technology.

It took another decade and the introduction of global standards for digital mobile systems to put a cellular phone in almost every person's hands. These systems basically provide the same service as previous analog systems, but employ advanced digital signal processing to lower the cost of production and to improve the range and the tolerance to interference, allowing more users in the system without compromising the speech quality. In particular, systems based on the global GSM (Global System for Mobile communications) standard had become immensely popular around the globe. Recent statistics in 2011 showed that there are more than 6 billion mobile subscriptions world wide, corresponding to an 87% penetration globally.

As new wireless systems evolved to complement the second-generation wireless access systems, a distinct shift in design criteria can be noted. From being primarily systems for voice communication, 3G wireless systems appearing on the market in the early 2000s were designed to also deal with data and various multimedia and web-based applications. The globally most popular 3G system, UMTS (Universal Mobile Telecommunications System), is not *per se* very much more efficient than its 2G predecessors but is geared to provide a combination of packet- and circuit-switched low-level bearer services, initially with on-air data rates between 384 kbit/s wide area coverage to 2 Mbps indoor/microcell coverage. Already a few years later high-speed data-only evolutions 3.5G or Turbo 3G were introduced in UMTS providing raw on-air data rates up to 14 Mbps with significantly lower latency, improving in particular the performance of web-originated downlink traffic.

Early 3G systems were based on the concept of generic bearer services, each mapping to some specific class of end-user applications envisaged by the system designers. In recent years it has become more and more obvious that it is impossible to predict what applications will be so popular, so-called killer applications, that we should design the wireless access infrastructure around them. Most 3G systems nowadays provide only two types of bearer services—voice and best-effort packet services—the latter to provide generic IP (Internet Protocol) access. One may say that IP access has become the completely dominant communication service—the killer service—both for fixed and for mobile applications. As a consequence, 4G systems, e.g. Long-Term Evolution (LTE) being deployed currently, are data-only systems providing wireless IP access with high data rates up to 100 Mbps and low latency. In parallel, we have seen the evolution of wireless local area networks, which are aiming to provide similar services but at much higher data rates at short ranges in office environments.

Fueled by the introduction of flat rate tariffing, the commercial success of mobile and wireless access to the Internet has been monumental in recent years. Initially thought of as a way of selling excess capacity in 3G networks, or providing some simple value-added services, it has, together with the proliferation of smartphones, created an explosion of traffic volumes. This trend is already threatening to overrun many networks. It is obvious that new technologies and careful management of resources in terms of cost, energy and radio frequency spectrum is needed to keep up with the pace of these developments. These problems are indeed the main focus of this book.

1.2 Key problems in wireless systems

The designers of wireless systems have struggled with a series of fundamental design bottlenecks, or key problems, each typical of their respective phase of development.

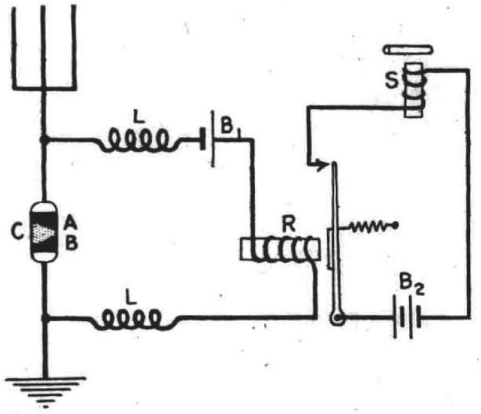


Figure 1.1 Early radio: Marconi-type passive coherer radio receivers. From *Elements of Radiotelegraphy*, E. W. Stone, 1919.

The removal of each bottleneck made the development take a significant leap forward, just to face another bottleneck. Let's briefly review these key problems.

1.2.1 Path loss—the early days

The dawn of wireless communication was dominated by wireless telegraphy, which became widespread in the early years of the 20th century. As electronic amplifiers were yet to be invented, the receivers were passive devices, mainly consisting of a simple tuned circuit, i.e. a bandpass filter tuned to the dominant frequency of the transmissions. As the signals could not be amplified in the receiver, all the energy at the receiver output (e.g. a sound in an earphone or the energy for pulling on a magnet that would operate a pen) had to be generated at the transmitter (see Figure 1.1). The loss of energy over a wireless connection, the path loss, is gigantic, in particular over large distances. This meant that the transmitters were large and bulky, capable of radiating enormous amounts of power. Needless to say, this was a severe limitation to mobile wireless communication, maybe with the exception of use on larger ships.

As the electronic tube amplifier [L. de Forest, 1907] became available, the path loss problem could be solved. Now, as receivers could amplify the weak received signals almost arbitrarily, the path loss could be completely compensated for. Moderately sized transmitters and sound transmission took us into the era of radio broadcasting in the early 1920s. The word "radio" was now synonymous with radio broadcasting to the man on the street. In the western world, a radio receiver appeared in virtually every home. Sound broadcasting was soon followed by TV broadcasting. In the USA this occurred in the 1930s, but elsewhere the commercial success of TV had to wait for the 1950s. Wireless communication also played a pivotal role in the Second World War. Soon after the war two groundbreaking inventions revolutionized wireless communication. The first was the invention of the transistor. Originally intended as a tube replacement in order to enable low-power, portable radios, the term “transistor” became synonymous with the small pocket broadcast receiver in the 1950s and 1960s.

At this stage, when there was now no limit to the amplification of signals in the receivers, communication engineers became aware of the next bottleneck, the thermal noise.

1.2.2 Thermal noise

Thermal noise is caused by the Brownian dance of electrons and is everywhere. It appears in all materials and electronic components. No matter how much the received signals are amplified, the noise will also be amplified alongside them. The second key discovery at the end of the Second World War was the recognition of the fundamental limits to the amount of information that could be transmitted reliably in the presence of noise.

As Claude E. Shannon published his “A mathematical theory of communication” in 1948, he laid the foundation of digital communications. At the time, the findings were not very practicable as the advent of integrated circuits and digital signal processing (DSP) devices was still decades away. We can now push the performance of today’s wireless communication systems very close to the Shannon limits. The most remarkable achievements made possible by this new way of thinking are probably satellite communication and communicating with deep space probes. From a commercial angle, probably the most revolutionary item is the digital cellular telephone. The latter managed to provide acceptable voice quality even in the most adverse environments, in moving cars or indoors. Over the years, digital communication engineers have been quite successful in pushing the performance close to the constraints manifested in the laws of physics and in Shannon’s theory. However, as we got more and more cellular phones, another fundamental problem became evident: the limited radio spectrum.

1.2.3 Interference—the limited spectrum

Although Shannon had already noted some aspects of this problem in the study of bandlimited channels, it has become evident that this is not entirely a technical problem. Since there is only one ether, it is clear that extensive and concurrent use of the same natural resource will inevitably lead to conflicts, in this case to unwanted interference between different users. This was obvious to radio listeners in the old days as they tried to receive a radio program on the medium wave AM band at night. Hundreds of radio stations competed for the attention of the listeners creating devastating mutual interference. The signal strength is in most cases sufficiently high that it would be possible to properly receive most of these stations if they were alone. This means that the problem is something different from the struggle against nature, i.e. the thermal noise discussed above. Rather, this problem, as with all resource-sharing problems, also has a social dimension. This dimension was already recognized in the very early years of radio, when the sharing of the frequency spectrum was given an administrative solution.

The International Telecommunication Union (ITU) was formed just after the Second World War specifically to deal with these problems. A technique that has been popular for spectrum resource sharing since the advent of radio communication is frequency multiplexing. The available spectrum is split into narrow frequency bands, mainly because early modulation schemes produced narrowband signals. This was an excellent way to separate different users of the spectrum and to avoid unintended interference. Within the ITU, the countries of the world have collaborated to closely regulate the use of the frequency spectrum. This spectrum hierarchy starts at the ITU where the frequency spectrum from 10 kHz to 200 GHz is meticulously split down into almost 100 bands or allocations. These allocations are in turn assigned to services in a document called the Radio Regulations (RR). Among these services you find fixed, mobile, broadcasting, radar, amateur and other similar uses of the radio spectrum. The frequency allocations are not at this level assigned to owners, nor to any country. Assigning spectrum to individual users is normally done by the National Regulatory Agencies (NRAs) in each of the ITU member countries. Frequencies are let to different users and user groups using various licensing arrangements. Licenses are typically issued for considerable periods of time, to match the technical/economical lifespan of the radio equipment used. The NRAs guarantee (police) that the provisions of the RR are maintained. When it comes to the lower frequency bands with long-distance propagation properties, decisions cannot be taken by individual NRAs, but instead all permissions for new transmitter sites have to be internationally coordinated. In principle, this requires that for every new transmitter, the NRA of that country has to collect the consent of all other countries that could be affected within some reasonable range. Needless to say, as the usage of the spectrum has been increasing, this has become a complicated matter. As new technical developments have meant new requirements on the spectrum, the rather slow administrative process has had difficulties in keeping up.

At the ITU level, making any significant changes in the RR, e.g. to allocate frequency bands to new systems and technologies, has been a demanding task, since a consensus decision between over 170 member states of the ITU has to be reached. Such changes are discussed at the World (Administrative) Radio Conferences (WARC or WRC) which are held every fourth or fifth year. One reason for this is the large differences in wealth and technological development in different parts of the world. Whereas some countries, e.g. western highly developed countries, demand new systems with high capacity and performance, other countries may have the view that the old technology is still viable and that already-made investments in equipment, receivers and so on should be protected. Major changes, if even possible to decide on, may require decades of careful planning and lobbying.

As the 1980s saw the transition from land mobile radio to (automated) mass-market mobile telephony, it became clear that a radically different solution was needed. It is obvious that individual users due to their sheer number cannot be given individual frequency assignments by the NRAs, and even more that the NRAs would be able to protect their reception quality. Instead, in a cellular telephone system the frequency administration is largely handed over to the owner of the system, the operator. The operator is given a license and frequency assignments by the NRA. When designing a system the operator has to organize the use of the spectrum in such a way that interference between the users of the system is kept to an acceptable level. Cellular telephone systems use a combination of careful planning and automatic schemes that adapt the spectrum utilization to the current user requirements. In the planning stage, the base stations are placed at carefully chosen locations and each base station is assigned a certain set of frequencies. The choice of which base station is to be used for connecting a mobile telephone and which actual frequency channel is to be used is done automatically while the system is in operation.

1.2.4 Infrastructure cost and energy consumption

As we move from mobile telephony to mobile data systems with a several orders-of-magnitude increase in capacity and data rates, we are not able to find enough spectrum to match. In addition, at very high data rates our systems will also again become limited by the thermal noise (basically the second key problem again). To some extent, we can counteract the latter problem with more transmit power at the expense of increased energy consumption and low battery lifetime in mobile devices. Another, more effective, method to increase data rate and capacity is to use more base stations, thereby effectively limiting the range of transmissions. The price we pay for this is more investment in infrastructure and equipment. Balancing the infrastructure cost, the energy consumption and the available spectrum is today the key problem in high-capacity wireless systems. This problem set will be the main focus of this book.

1.3 Wireless access networks—the issues

There are many different types of wireless communication systems, ranging from broadcasting to satellite systems, from shipborne systems to police, fire brigade and military systems. In this book our focus will be on systems for mobile and nomadic (data) network access. The reason for this is that almost all IT services today rely on network access—information is retrieved, stored or processed remotely rather than in the mobile devices themselves. IP connectivity is becoming the dominant design in service provisioning, commercially dwarfing other types of specialized network solutions (e.g. peer-to-peer systems). We communicate with other people in distant locations and most of the apps in our smartphones are becoming cloud-based. Cloud computing is a consequence of efficient and virtually free communication. We compute and store information wherever in the world it may be cheapest and most effective, neglecting the cost of communication. The physical terminal we are using is of no consequence. In mobile access this has not really been the case—poor coverage, high cost, latency and limited data rates have been limitations that have prevented service mobility and convergence. If and when the mobile and wireless networks provide better connectivity and access speeds, cloud computing will inevitably also become the ruling paradigm in wireless. A striking example of this is when two friends meet in the street and would like to exchange digital content, say photos. In principle, short-range

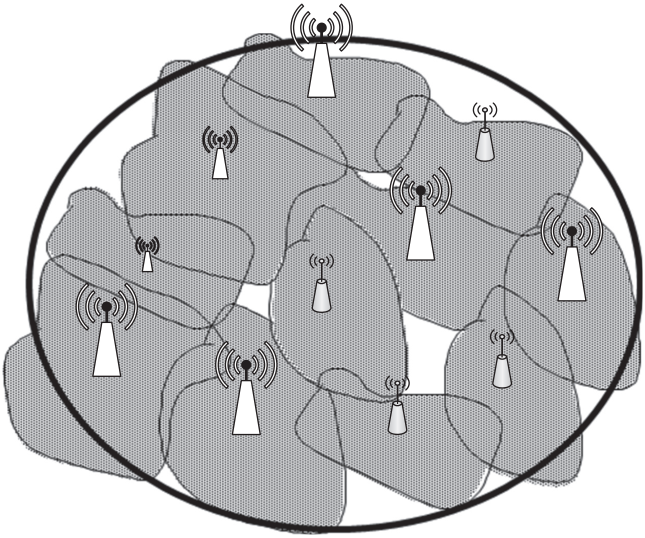


Figure 1.2 Schematic coverage map of a wireless communication system.

peer-to-peer radio connectivity, e.g. Bluetooth, would be the most effective way from an engineering perspective. However, instead of wasting time and effort in peering, you simply email your photo to your friend or put it on Flickr using cellular or WiFi access. The reason for this is that, even though this operation may consume significantly more network resources, the marginal cost for the user is zero.

In a wireless access system, the primary goal is to provide fixed network access to a large number of mobile or stationary users dispersed over a geographical area.

The number of users, their service demand and locations are not *a priori* known. An example from the early days of history is the (national) radio/TV broadcasting systems. In these systems, one-way wireless connections to individual mobile or stationary listeners are provided by a collection of broadcast transmitters connected to a program distribution network. Another, more recent example, which we will cover in somewhat more depth, is a mobile (cellular) telephone or mobile data system. In this example, the fixed infrastructure that the mobile users are attempting to acquire services from is the Public Switched Telephone Network (PSTN) or the Internet. To provide the services of the network, i.e. to connect mobile users to fixed (or other mobile) users in the network or servers, the network is extended by a set of radio base stations. The base stations provide the physical two-way radio connections to the mobile terminals. Figure 1.2 illustrates the principles of wireless network design. The network consists of a fixed network part and a wireless network part. The fixed network provides connections between base stations or Access Points (APs) which in turn provide the wireless connections to the mobiles. The APs are distributed over the geographical area where mobile users are provided with communication services.

This area is called the service area. The mobile terminals that are to be provided with the required service may be anywhere within the service area and will be assigned a connection to some AP. The assignment is done by the system and without any user intervention. The area around an AP where the transmission conditions are favorable enough to maintain a connection or provide a service of the required quality is termed the coverage area of the AP. The transmission quality (e.g. the voice quality, the data rate, etc.) and thus the shape of these regions will depend significantly on the propagation conditions and the interference from other users in the system.

The coverage areas of the individual APs are, in practice, of highly irregular shape. The areas may contain coverage holes, i.e. there are locations close to an AP that are not covered, e.g. due to shadowing from buildings. It is more common that the system is designed to create overlap areas, i.e. there are areas where a terminal may communicate with several APs. Also the opposite situation may occur, i.e. a situation where the terminal is in a white spot, a region where communication with sufficient quality is not possible. The fraction of the service area that is not affected by white spots is called the coverage or the area availability of the system. These quantities are both defined as the probability that communication is maintained at some given randomly chosen location (chosen from a uniform distribution) in the service area. Another measure of great interest is the population availability, i.e. the probability that a randomly selected user can be provided with adequate communication service. This measure can be calculated by weighting the covered areas with the (user) population density.

In two-way communication systems (such as mobile telephone or mobile IP access systems) links have to be established both from the AP to the mobile (called the downlink or forward link) and between the mobile terminal and the AP (the uplink or reverse link).

The first casual look may suggest that these links have very similar properties. There are, however, distinct differences from a radio propagation perspective. For example, in wide area cellular systems, the AP (base station) usually has its antennas at highly elevated locations, free from obstacles. The terminals, on the other hand, are usually located at street level, where buildings and other obstacles create shadowing and multipath reflections. Also the interference situation in the up- and downlinks will be different since there are many terminals with varying locations and relatively few APs at fixed locations.

For obvious economic reasons, a network owner wants to provide the required service at minimum cost. His/her objective will therefore be to provide sufficient coverage with as few APs as possible. This would not only minimize the cost of installation, towers, radio equipment and other AP hardware, but also minimize the fixed, wired part of the infrastructure. Various propagation effects limit the coverage and will thus put a lower limit on the number of APs that need to be installed. If the distance between two APs becomes too large, eventually there will be points between the APs where the signal level will drop too low, which will in turn result in poor voice quality or low data rate. Shadowing and multipath phenomena will add to these problems. Stated simply, the transmission range of the APs is too small compared to the inter-AP distance. Such a system where this type of problem is dominant is termed a range-limited system. Typically, mobile cellular systems are range-limited systems in their initial stages of development when the key objective is to quickly and at a low cost cover the service area for a low number of subscribers. Other examples are early broadcasting systems, where radio stations were few compared to the bandwidth available.

As systems evolve and become popular, cellular and broadcasting systems alike, the number of transmitters in the system eventually becomes large compared to the available bandwidth. These system are not primarily troubled by weak received signals, but interference from other APs and mobile terminals. Such systems are said to be bandwidth or interference limited. The main problem in these systems is the proper management of the scarce resources, e.g. bandwidth. The objective of this management task is to satisfy both the provider of services (the operator) and the user of these communication services. The former wants a high and efficient utilization of the system since he/she derives more revenues by providing higher data rates or services to more users, or he/she is capable of providing a given service with less resource consumption (less power, spectrum or APs). The user expects good Quality of Service (QoS). In cellular systems, such user requirements can be expressed in terms of probabilistic measures such as the average data rate, the probability of being denied making a voice call with acceptable quality (blocking) or when dropping an ongoing connection. In the following chapters, we will demonstrate that as in most resource management problems, the operators aim to increase the data rate or the number of users served. Such quantities we will loosely refer to as the capacity of the system. These objectives are in conflict with the users' desire to achieve a higher service quality. Squeezing more users into the system will inevitably cause more interference resulting in poorer transmission quality, lower data rates and/or longer waiting times. Striking the proper balance between these aims is a delicate problem for the operator when he/she makes offers to the users, in particular in a competitive situation. Efficient frequency resource management, i.e. employing schemes that either avoid some of the interference or that better resist the interference between users, can both increase the capacity and improve the service quality in the system.

The frequency spectrum is not the only resource wireless operators and their customers have to be concerned with—there are other scarce resources as well. One obvious such resource is the infrastructure of APs, including networks of switches/routers. It will become clear from our analysis that a denser system with more access ports (i.e. more expensive infrastructure), has the potential of providing more capacity and higher QoS to the users. Another important resource to be managed is the energy consumption in the system. Since most modern wireless networks are designed for lightweight and portable use, the battery energy is severely limited. Moreover, the biomedical restrictions on emitting electromagnetic fields from handheld devices also impose limits on the transmitter power. Limitations in available energy at the portable terminal may also lead to restrictions in the complexity of the signal processing algorithms employed at the terminal. In all these cases, lower transmitter power leads to either lower transmission quality or lower radio range. Each of these effects has to be countered by adding more access ports (i.e. a more expensive infrastructure). On the other hand, in mobile data systems with high data rate the cost of energy in the APs has also become a significant concern. In a similar way to trading off power requirements and infrastructure density, it will be seen that frequency spectrum bandwidth and power

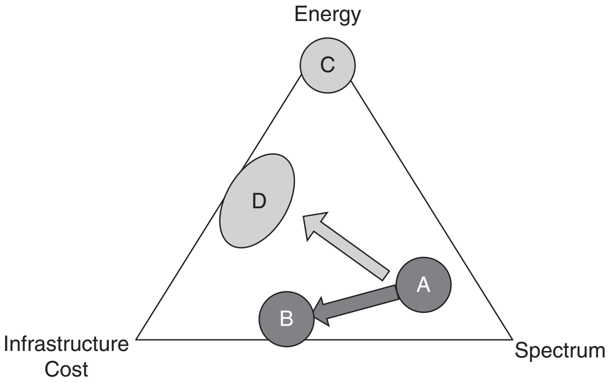


Figure 1.3 Trade-off of resources in wireless networks.

can be traded off. Figure 1.3 illustrates this interdependency. It illustrates how the traditional design of mobile communication systems has been spectrum limited (A). This situation is studied in Chapter 4. As systems require more capacity, wireless transmitters are packed closer and closer and the infrastructure cost (Chapter 11) starts to limit the design (B). As the cost for energy goes up, we have to be aware of this element as well. Completely energy-minimizing systems (C) we will see are not really realistic but a reasonable compromise has to be struck (D; Chapter 9).

1.4 Outline of the book

This book is intended as a textbook for an advanced course in wireless networks. We will assume that the students know the fundamentals of radio propagation and digital communications over wireless channels. Chapter 2 provides a more stringent definition of those models and performance metrics that were introduced in a more hand-waving fashion in Section 1.3 above. We also outline the basic methodology used to analyze wireless access networks. After this modeling introduction, the book is then basically divided into two parts.

Part I, Radio Resource Management (RRM) in wireless systems, discusses various techniques to manage the interference and to maximize the capacity in an existing (already deployed) wireless access network, first from an orthogonal access perspective and then a non-orthogonal access perspective. Chapter 3 discusses various medium access schemes and Chapter 4 various scheduling approaches. In both chapters, we assume orthogonal access and that no simultaneous transmissions are allowed on the same radio resources because of heavy interference between different users. Chapters 6 to 8 then discuss more advanced RRM techniques such as power control, interference management, handover, and other inter-cell interference management techniques. Chapter 11 illustrates how various RRM techniques are applied in 4G Long Term Evolution (LTE) systems.

Part II, deployment of wireless access networks, studies how to dimension and deploy wireless networks. In Chapter 5 we introduce basic cellular system concepts and analyze cellular network capacity for networks with either orthogonal or non-orthogonal medium access implemented in each cell. We will introduce the important concepts of range- and interference-limited systems and study the statistical properties of inter-cell interference. Some basic relations between the coverage, available spectrum, AP density, medium access and the capacity of the systems will be analyzed. This chapter will also describe how to dimension voice and data networks to meet traffic demands. Heterogeneous cellular structure and antenna techniques are also discussed. Chapter 9 discusses the impact of energy consumption on wireless network operations and the optimal energy-efficient design. Finally, Chapter 11 provides the fundamentals of wireless infrastructure economics, also taking various cost items into account to achieve a cost-efficient network deployment.

References

L. de Forest. 1907. Device for amplifying feeble electrical currents. US Patent 841387 A.

2 Wireless network models

2.1 Introduction

Looking at classical communication theory, we see that it mainly deals with point-to-point links disturbed by thermal (Gaussian) noise. More recently the challenges of mobile communication have introduced features such as adverse, time-varying propagation conditions which create channel variations that are difficult to predict. Radio systems, as we find them in reality, have to cope with additional problems. Maybe the most characteristic feature of modern radio communication is that virtually no radio link or radio system operates in isolation, and is thus never alone in its allocated frequency band. Other radio transmitters, near and far, constantly cause interference. Interference is in many cases the limiting factor to the performance of the system. Since the days of Marconi, the proliferation of wireless communications has caused a tremendous increase in the utilization of the frequency spectrum. A key problem area, as was already noted in Chapter 1, is how to manage the frequency spectrum to avoid, or at least minimize, the adverse effects of interference. Can interference be avoided completely, or are there efficient methods to minimize the performance degradation? The ether, where we transmit our signals, is, whether intended or not, a broadcast medium. In some geographical regions, a large number of wireless networks have to coexist as illustrated in Figure 2.1. The blessing of wireless communication is that it allows for quickly establishing arbitrary new connections between a large number of users. In Figure 2.1, we consider three transmitters transmitting information to three different receivers indicated by the solid black arrows. We call these paths the active communication links. As the radio spectrum is shared by all users the transmissions of the three transmitters in Figure 2.1 give rise to interference. These unwanted cross-links, the interference links, are indicated by the light gray arrows in the figure.

The properties of the interference will depend on the waveforms and transmitter powers selected by the interfering transmitters as well as the propagation conditions on the cross-links. The impact on the performance of the active communication link will depend not only on the waveforms, the powers and the propagation conditions in the active link but also on the performance of the radio receiver, e.g. how good the receiver is at suppressing the unwanted signals. The performance experienced by the user in this network will depend on the type of service that is provided. In a mobile phone network it may be the voice quality that is important; in a mobile data network,

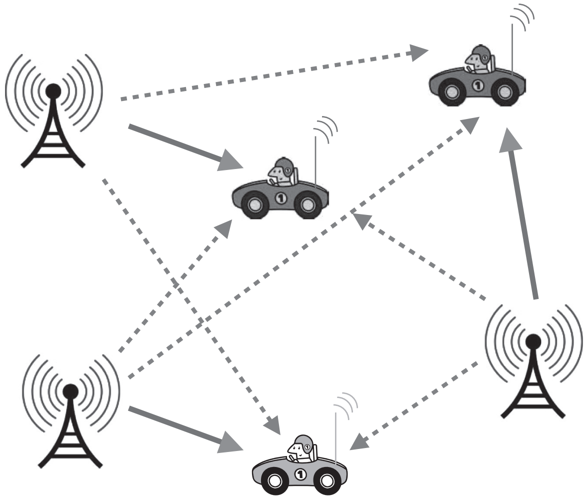


Figure 2.1 A wireless network.

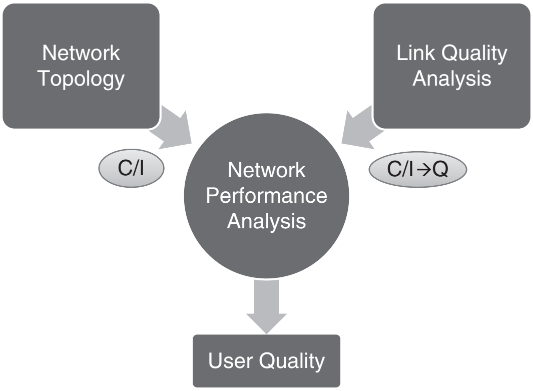


Figure 2.2 Simplified network analysis strategy.

the user will be interested in the download time of a web page or that a video is played out smoothly on their smartphone screen without interruptions.

A practical network with thousands of wireless transmitters and receivers with millions of potential interactions makes a detailed analysis a very demanding, or even impractical, task. In order to get a general understanding of the behavior of such a network we need to make some approximations to simplify the analysis. The most common analysis approach is to use a two-step procedure as illustrated in Figure 2.2. We will follow this procedure throughout this book. As a first step, we characterize the interference by its power as experienced at the receiver. This involves modeling the propagation path losses which, in turn, are dependent on the relative locations of the transmitters and receivers and the specific propagation conditions determined by the terrain in between.

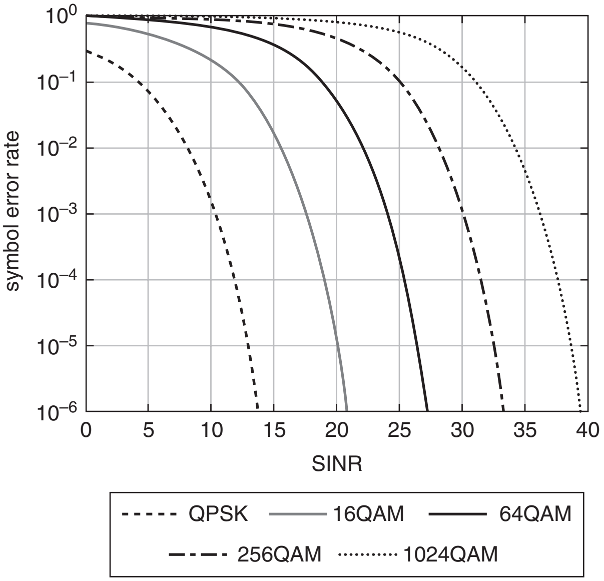


Figure 2.3 Symbol error rate as a function of signal to noise ratio (SNR) for four different modulation/coding schemes.

In the second step the impact of that interference on the performance of the active link is analyzed. For this purpose we consider the Signal to Interference plus Noise Ratio (SINR) as our basic performance measure. The instantaneous SINR is only a function of the network topology and the propagation conditions. It is dependent on the user activity, e.g. if a terminal is transmitting with a certain power or not, but not explicitly on the user service. This is a quantity that can, at least in principle, be objectively measured at the receiver.

Armed with this rough quality measure, we will now derive an approximate mapping of the average SINR to the user-perceived link quality performance. As our simple model disregards the finer structure of the received waveform, which is also affected by the propagation conditions, e.g. multipath delay spread, doppler shift conditions and so forth, what we can hope for is an estimate of the average user-perceived performance. Figure 2.3 shows an example of how such a mapping may work [P. Schramm et al., 1998] for a GSM data (GPRS, General Packet Radio Service) link. Here we first estimate the symbol error rate as a function of the SINR, and then this error rate is used to estimate the average data rate or download time of a web page for the individual user. We will discuss such measures in more detail in the next section.

2.2 Models for wireless access networks

We will now apply the above methodology to a mobile wireless access network as outlined in Chapter 1. A mobile access network is a wireless communication system where mobile users (terminals) move around in the service area and will from time to time require communication services in the form of a wireless connection or packet transmission to and from a fixed network. As previously noted, like most resource allocation problems this one has two distinct aspects, a provider perspective and a consumer perspective. The network provider, i.e. the operator, owns and provides the communication resources to the users who consume them. The two parties have different interests. The operator wishes to run a profitable business, whereas the user is willing to pay for the quality services he/she appreciates. In the simplest case where all users are provided with the same service offering for the same cost, the revenues of the operator will be maximized if he/she can maximize the number of users using the system. The latter certainly depends on economic factors such as the price and competing operators or services, but also on the technical limitations of the systems.

This capability of a network to handle many users, transferring many bits, is usually loosely referred to as the capacity of the network. These measures should not be confused with the information theoretic concept with the same name. Depending on the character of the service provided, more precise definitions of the capacity are possible. In a simple scenario with a single service and constant data rate, e.g. mobile telephony, the capacity may be defined as the maximum number of users that can be served by the system. In a mobile data scenario, we could also define the capacity as the maximal total data rate transferred by the system. The user in turn is interested in getting the best QoS, which could be good voice quality, high data rates or low delay. It is obvious that services requiring more resources per user will limit the capacity of the system. In general, the capacity will be a function of these QoS requirements. Since the operator is paying for the system, it is natural that the design process is ruled by his/her perspective. The problem will thus be to design systems where the profit, i.e. revenue – cost, is maximized for some given service requirement of the user. If the latter requirement is not met, the user will not use the service and no revenue can be derived. In general, however, a complication is that perceived service quality is not easily related to the willingness of the user to pay and the resources consumed in the network. We will discuss this in more detail in Chapter 11. For the time being, we will instead formulate a somewhat simpler problem in purely engineering terms. We assume that the operator seeks to maximize the system capacity for a given technical measurable QoS requirement. The latter will be the constraint in our optimization problem. In this chapter we will provide a more formal definition of this problem for some service cases. Designing a wireless access system involves two different tasks:

- *The network deployment problem:* Here the problem of designing the network infrastructure is considered. How many APs are required? Where should they be placed? What fixed network capacity has to be provided for the different APs? How much spectrum should be allocated to the system? Since this is a long-term process, business decisions regarding deployment and dimensioning of the network have to be based on long-term predictions/estimates of the user demand.
- *The radio resource management problem:* Given a certain infrastructure deployment, i.e. AP locations, spectrum allocation, fixed network capacity and so on, how

should the wireless resources be allocated to maximize capacity, to best meet the instantaneous demand of the users/mobile terminals moving around in the network? This is the problem addressed in more depth in this chapter.

These two tasks are of course intimately related. In order to make decisions regarding the deployment of certain configurations of APs, we need to know what performance can be achieved by the radio resource management schemes in these configurations. This is why we will start by studying the radio resource management problem in this and the next few chapters. We will return to the deployment problem in Chapters 5 and 11.

2.3 Service scenarios and performance metrics

We need to understand that the user is interested in receiving a communication or information service in a wider sense than in the technical systems that are studied here. Whereas the systems we consider in this book transport bits and messages, they are only a part of a complex service chain. Mapping perceived performance to technical parameters in the wireless system is therefore a very complex task. The user experience is influenced not only by the shortcomings of the wireless system but also by other factors such as the performance of the wired backbone and switching, the service provider's application software and hardware, and the user interface provided by the service provider and the terminal manufacturer (to give a few examples). In order to allow rational design of wireless access networks, these overall end-to-end requirements are broken down to specific service requirements for the individual building blocks. Here the interest is focused on studying the behavior of the radio network part of the transport system. Some examples of the technical quality of service parameters that could be used to characterize such a transport service are:

- *Maximum data rate*: The highest data rate averaged over a certain time interval that a user could expect.
- *Guaranteed data rate*: The lowest data rate averaged over a certain time interval that a user is guaranteed.
- *Residual bit error rate*: The undetected error rate after the delivery of the information over the service interface.
- *Transmission delay*: The time delay the packet/message spends between the service access points.

There are many ways in which such QoS requirements can be met in a mobile communication system. In the UMTS/3G standardization process in the late 1990s and early 2000s, the ruling paradigm was to solve QoS problems at the radio network level. Significant efforts were made to design so-called bearer services [3GPP, 1999]. Table 2.1 illustrates the four classes of bearers. In each class, specific sets of service parameters, forming a QoS profile, would be used to match the end-user services and to maximize the perceived QoS. However, as we noted in Chapter 1, the victory march

Table 2.1 3G (UMTS) bearer service classes [3GPP, 1999]

Service class	Typical applications	Service functional characteristics
Conversational real time (RT)	Voice	• Preserve time relations between entities • Stringent preservation of conversational patterns with low delay
Streaming RT	Video/audio streams	Preserve time relations between entities
Interactive best effort (BE)	Web browsing	• Request–response pattern • Preserve payload (low error rate)
Background BE	File transfer, e-mail	• Not time critical • Preserve payload (low error rate)

of IP has made these efforts literally collapse. The wireless system providers are, with the exception of voice and short message services (SMS), basically ignorant of what new and ever-changing services are provided over the top of their systems using the IP layer. The consequence is that 3G systems nowadays in practice only use two types of standardized bearers, a circuit-switched bearer service for voice and a best-effort service for IP traffic. In 4G (High Speed Packet Access, LTE) systems, the voice bearer is omitted as the capacity requirements for the voice services are dwarfed by the demand from other services. The key reason for this trend is our inability to foresee what new services will appear and what QoS they will require.

Over-providing best-effort resources is the brute-force solution to most QoS problems, present and future. One can say that in the IP world we have always been prepared to waste a lot of precious resources of cost, energy and spectrum to achieve transparency, i.e. that our network is capable of handling all kinds of services, even those that haven't been invented yet. This choice has been easy in fixed networks where optical fiber technology from the mid-1990s has provided vast amounts of capacity. In the wireless domain this approach has taken more time to realize, but the advent of the iPhone in 2007 can be said to mark this paradigm shift.

In this chapter we will therefore limit our analysis to two simple models for user behavior and QoS requirements:

Guaranteed service quality—blocking: the user will request a service that has to achieve at least a certain minimum QoS. If this service level cannot be achieved for some users, the system will opt not to commit any resources to these users. Some of the users will be blocked, and will not be provided any service. A typical example of such a service is voice. Once the QoS target is reached, it is meaningless to provide a better connection. On the other hand, a modern digital voice connection deteriorates rapidly once we drop below the minimum link quality (Figure 2.4).

Best effort—non-blocking: In this model the perceived quality for the user will monotonically increase. The system will provide the service to all the active users. A typical example is mobile data. All active users share the available resources and no user is blocked.

As discussed in the previous section, we will make the simplification that the QoS is completely determined by the link SINR.

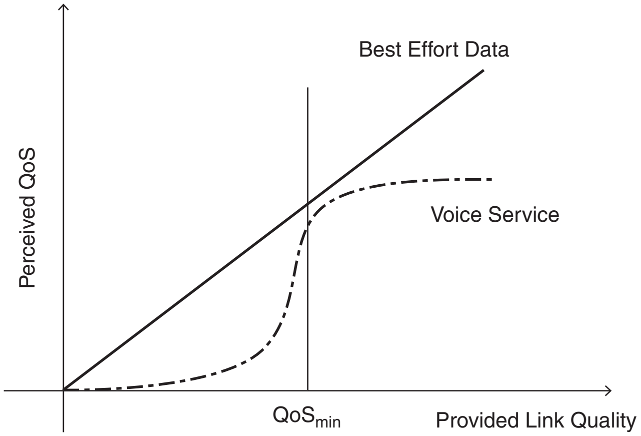


Figure 2.4 Perceived quality of service for the two service models.

2.4 Radio resource management in wireless access networks

Now assume that we have deployed a certain number of APs in our service area (Figure 2.5). In the service area, mobile terminals will roam. A mobile terminal requiring services, e.g. a connection, will be called an active terminal. The system will, whenever possible, attempt to establish a physical connection between an active terminal and some access port in the network. Assume that, at a certain instant, there are M active terminals in the service area numbered using the set

$$\boldsymbol{M} = \{1, 2, 3, \ldots, M\}. \tag{2.1}$$

The number of active terminals will fluctuate as time goes by. This means that also the size of this set $\boldsymbol{M}$ will vary. We will model $\boldsymbol{M}$ as a stochastic variable. The distribution of $\boldsymbol{M}$ will depend on user behavior. Parameters such as the traffic intensity and the duration of a user session will come into play. In addition, the distribution will also depend on the system behavior, e.g. to what extent the service requirements of the individual users can be satisfied. Users will be assumed to initiate and terminate sessions according to some random processes, which significantly complicates our analysis. The problem is approached in several ways. Mostly, throughout the book, we will try to establish how these services can be provided to the users at some instant of time. This is referred to as a snapshot analysis. This type of approach is of direct relevance to all kinds of circuit-switched service, e.g. voice, or virtual circuit connections, e.g. downloading large files, streaming video and so on, where the users require physical access to the radio medium at almost all times. This means that we will need to fulfill the SINR requirements at virtually any snapshot, or at least on the average over a number of subsequent snapshots. Non-real-time, less delay-sensitive services like email, messaging and so on will be treated in more detail in Chapter 3. In the following snapshot analysis, we assume that the active terminals are *uniformly distributed* over the service area. This means that the probability density of the location $R = (X, Y)$ of

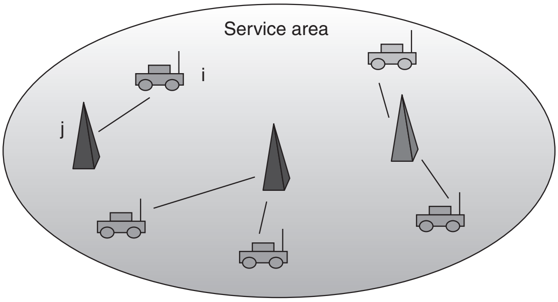


Figure 2.5 The radio resource management problem.

some given active terminal is constant,

$$\begin{aligned} p(x,y)dxdy &= \Pr[X \in [x,x+dx], Y \in [y,y+dy]] \\ &= \frac{1}{A}dxdy, \end{aligned} \tag{2.2}$$

where A is the area of the service region (see the definition in Chapter 1). A certain terminal is thus equally likely to be found in any part of the service region. If the number of active terminals is large and if the terminals are activated independently of each other, point process theory tells us that the locations of the active terminals form a two-dimensional Poisson point process. A property of two-dimensional Poisson processes is that the numbers of points in disjoint areas are independent and Poisson-distributed random variables. Given that a certain number of active terminals are observed in some given area, the locations of these terminals will be independent and uniformly distributed over that particular area. Let the intensity (rate) of this process be denoted by ω active terminals/area unit. It is obvious that the individual terminals do not act quite independently, as they share or compete for the same radio resources. Still, this independence assumption may serve as a reasonable approximation if the traffic load is not too high, i.e. when most terminals are assigned sufficient resources. The relationship between the frequency of session starts, their duration and the area A on one side and ω on the other is not trivial and will depend on the behavior of the users as well as on the design of the system.

The terminals are associated with and served by access ports or base stations, numbered from the set

$$\boldsymbol{B} = \{1,2,3,\ldots,N\}. \tag{2.3}$$

Now assume that there are C waveforms (channels) available for establishing links between access ports and terminals. We number these from the set

$$\boldsymbol{C} = \{1,2,3,\ldots,C\}. \tag{2.4}$$

In order to establish wireless connections between access ports and terminals, the system is required to assign to each terminal

a) an access port from the set $\boldsymbol{B}$,
b) a waveform (channel) from the set $\boldsymbol{C}$,
c) a transmitter power for the access port and terminal station.

The way we assign access port, channel and power is done according to the *resource allocation algorithm* (RAA) of our system. The assignment is limited by the *thermal noise* if terminals are far from the AP and by the interference caused by other access ports and terminals. Another common constraint is that access ports are sometimes only capable of using a certain subset of the available waveforms. Good allocation schemes will aim at assigning links with adequate SINR to as many terminals as possible. Note that the RAA may well choose not to assign a waveform to an active terminal if this assignment would cause too much interference to other terminals.

Example 2.1: In a mobile telephone system, terminals in a certain area are assigned one access port. The access port has in turn been assigned N channels. Calls with exponentially distributed holding times and Poisson arrivals are blocked if all N channels are busy. Show that if N is large, the number of calls in progress in the area in a certain snapshot is approximately Poisson-distributed.

Solution: Let $n(t)$ denote the number of calls at time t. $n(t)$ is said to be in state k if k channels are busy, which is the same as having k calls in progress. It is easy to see that $n(t)$ is a birth–death process with rates

$$\lambda_k = \begin{cases} \lambda & \text{if } k < N \\ 0 & \text{else} \end{cases} \tag{2.5}$$

and

$$\mu_k = k\mu, \qquad k = 1, 2, \ldots, N. \tag{2.6}$$

The state transition diagram is illustrated in Figure 2.6. The steady-state distribution exists as there is a finite state space. The distribution can be obtained as

$$P_k = \begin{cases} P_0 \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!} & \text{if } k \le N \\ 0 & \text{else.} \end{cases} \tag{2.7}$$

Since $\sum_k P_k = 1$, we have

$$P_0 = \left(\sum_{k=0}^{n} \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!} \right)^{-1} \tag{2.8}$$

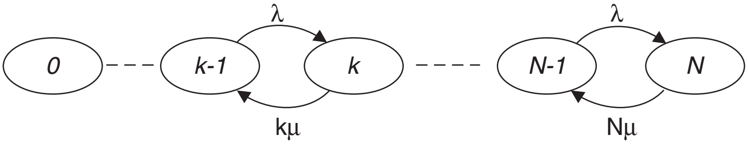


Figure 2.6 State transition.

and letting $\rho = \frac{\lambda}{\mu}$, for $k \leq N$

$$P_k = \frac{\left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!}}{\sum_{j=0}^{n} \left(\frac{\lambda}{\mu}\right)^j \frac{1}{j!}} = \frac{\frac{\rho^k}{k!}}{\sum_{j=0}^{n} \frac{\rho^j}{j!}}, \tag{2.9}$$

which is referred to as Erlang's B-formula. As $N \to \infty$,

$$\lim_{N\to\infty} p_k = \frac{\rho^k}{k!} e^{-\rho}, \tag{2.10}$$

which is a Poisson distribution.

The analysis approach outlined in the previous chapter showed that only the signal and interference power levels at all access ports and terminals have to be computed. We can do this using the link gains between transmitters and receivers, i.e. where g_{ij} is the power gain experienced by the signal on the path between access port i and terminal j. This, in turn, means that the received power in receiver j, denoted by $P_{rx,j}$, can be computed as

$$P_{rx,j} = P_{tx,i} g_{ij}, \tag{2.11}$$

where $P_{tx,i}$ is the transmitter power assigned to transmitter i. In the following we will treat all the g_{ij}'s as random variables. Putting all these link gains in a matrix results in a $B \times M$ rectangular matrix, the *link power gain matrix*:

$$\boldsymbol{G} = \begin{pmatrix} g_{11} & g_{12} & \cdots & \cdots & g_{1M} \\ g_{21} & g_{22} & \cdots & \cdots & g_{2M} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ g_{B1} & g_{B2} & \cdots & \cdots & g_{BM} \end{pmatrix} \tag{2.12}$$

The link gain matrix can be said to describe all the instantaneous propagation conditions in the entire system. It is important to note that in a practical system the number of active terminals varies and the terminals move. This means that both the individual components and the dimension of the matrix may vary over time.

A resource allocation scheme should find an assignment for which a QoS can be achieved in as many links as possible. In our snapshot analysis, the instantaneous transmission quality should therefore be sufficient. This in turn means that the signal to interference ratios (SIRs), or actually the SINRs, are larger than some requirement. This requirement could possibly be different for different links. Given that a terminal j has been assigned to access port i_o using waveform $c(i_o)$, the SINR in the uplink and downlink of the connection can be computed as

$$\Gamma^u_{i_0,j} = \frac{P_j g_{i_0 j}}{\sum_{m \neq j} P_m \kappa_{c(i_0),c(m)} g_{i_0 m} + N_{i_0}} \tag{2.13}$$

$$\Gamma^d_{i_0,j} = \frac{P_{i_0} g_{i_0,j}}{\sum_{b \neq i_0} P_b \kappa_{c(i_0),c(b)} g_{bj} + N_j}, \tag{2.14}$$

where $\Gamma^u_{i_0j}, \Gamma^d_{i_0j}$ denote the uplink and downlink SINR respectively. N_j and N_{i_0} denote the receiver thermal noise power at the terminal and access port respectively. κ_{ij} denotes the cross-correlations between waveforms i and j. The effective interference power received at terminal m from access port b is thus $P_b\kappa_{m,b}$. The total interference power is assumed to be the sum of these effective powers of the individual interference components.

2.4.1 Orthogonal signal sets

An important special case occurs when $\boldsymbol{C}$ consists of orthogonal waveforms only, i.e.

$$\kappa_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases} \tag{2.15}$$

Waveform sets of this type are found in first generation mobile telephony systems. Here the waveforms used are the historically very popular *FDMA waveforms*, i.e. non-overlapping narrowband signals or frequency channels (Frequency Division Multiple Access). In this case we can refer to the assignment problem as the *channel allocation* problem. The SINR expressions (2.13), (2.14) in this case simplify to

$$\Gamma^u_{i_0j} = \frac{P_j g_{i_0j}}{\sum_{m\neq j, m\in M^{(c)}} P_m g_{i_0m} + N_{i_0}} \tag{2.16}$$

$$\Gamma^d_{i_0j} = \frac{P_{i_0} g_{i_0j}}{\sum_{b\neq i_0, b\in B^{(c)}} P_b g_{bj} + N_j}, \tag{2.17}$$

where the interference in this case has to be added up over all terminals and APs using the same channel c only.

It is clear that there is no way to guarantee that we will comply with all the constraints for all the M terminals, in particular for large M and B. We may have to be satisfied with finding resource allocation schemes that assign waveforms with reasonable quality to as many terminals as possible. The precise definition of our performance measure, in the previous chapter loosely referred to as the system capacity, will depend on which of the specific service cases we study. Remember that the number of active calls is the random number M and that the link gains in the matrix $\boldsymbol{G}$ can be modeled as stochastic variables. For our two archetypical service types, we have the following considerations.

2.4.2 Guaranteed service quality—blocking

Assume that at some given instant our RAA has succeeded in providing waveforms that provide adequate quality. There are Y of these terminals for which

$$\Gamma^u_{i_0,j} = \frac{P_j g_{i_0j}}{\sum_{m\neq j} P_m \kappa_{0,m} g_{i_0m} + N_{i_0}} \geq \gamma^u_j, \tag{2.18}$$

$$\Gamma^d_{i_0,j} = \frac{P_{i_0} g_{i_0,j}}{\sum_{b\neq i_0} P_b \kappa_{0,b} g_{bj} + N_j} \geq \gamma^d_j \tag{2.19}$$

are the required SINRs in the up- and downlinks for the required service. Y will of course also be a stochastic variable. Let us denote by Z the remaining number of terminals that have not been assigned a channel, the number of *assignment failures*:

$$Z = M - Y. \tag{2.20}$$

The *assignment failure rate* ν can now be defined as

$$\nu = \frac{E\{Z\}}{E\{M\}} = \frac{E\{Z\}}{\omega A}. \tag{2.21}$$

This quantity is a measure of how successful the allocation scheme has been, on average, in providing the terminals with links of adequate quality. For moderate to large ωA, ν is also a good approximation of the probability that some arbitrarily chosen active terminal is not provided with the proper combination of waveforms and access ports without violating the constraints (2.2). The instantaneous capacity $\omega^*(\nu_0)$ of a wireless system is the maximum traffic load that can be permitted that still keeps the assignment failure rate below some threshold level ν_0, i.e.

$$\omega^*(\nu_0) = \{\max \omega : \nu \leq \nu_0\}. \tag{2.22}$$

The capacity can therefore be expressed in terms of the number of active terminals/area unit.

2.4.3 Best effort—non-blocking

In this service scenario we assume that the performance metric is the data rate in the link to mobile i as a function of the SINR,

$$R_i = f(\Gamma_i). \tag{2.23}$$

The capacity is then defined as

$$\bar{R}^* = E\left\{\sum_{i=1}^{M} R_i\right\} = E\left\{\sum_{i=1}^{M} f(\Gamma_i)\right\}, \tag{2.24}$$

the total sum rate that can be delivered by the system in the service area. A popular model for modern mobile data systems with modulation and coding schemes operating very close to the information theoretical limits is to use the Shannon bound to describe the relationship between rate and SINR in a channel with bandwidth W:

$$R_i = \min\left(R_{\max}, cW \log_2 (1 + \Gamma_i)\right), \tag{2.25}$$

where $R_{\max}$ denotes the maximum data rate that the terminal/AP hardware is capable of delivering and the constant c is a function of implementation imperfections.

Resource allocation algorithms can be designed in many ways for either case. Figure 2.7 illustrates how an RAA uses information through measurements about the traffic to make its decisions about the access port association, the waveforms and the transmitter power assignment. This information includes the number of active users,

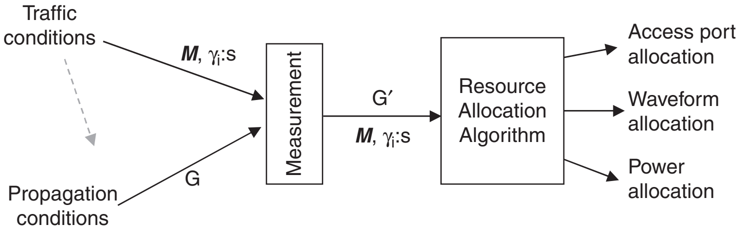


Figure 2.7 Resource allocation algorithm operation.

their QoS requirements and the propagation conditions (i.e. link gain matrix $\boldsymbol{G}$). The discussion above has shown that finding the optimum resource allocation, i.e. the waveform, the power and the access port assignments that maximize Y or finding the maximum $\bar{R}^*$ for a given user distribution and corresponding link gain matrix, is a formidable problem. We can actually show that this problem in this general form belongs to the set of NP-complete problems. For this particular class of problems, we know of no efficient general algorithm that is capable of solving for an optimal resource assignment for arbitrary link gain matrices and arbitrary sets of access points and terminals. By "efficient" we mean an algorithm where the number of computational steps does not increase exponentially with the size of the problem, e.g. the number of APs and terminals. In practical systems, the focus has instead been on more simple, low-complexity heuristic schemes that are used in many of today's wireless systems. The capacity ω^* or $\bar{R}^*$ achieved by these schemes is, as expected, often somewhat lower than what could have been achieved by optimum waveform/access port/power assignment. Surprisingly, many of these simplified schemes achieve a performance that is not far from the optimum.

In our discussion so far we have considered a specific snapshot. As time passes, propagation conditions change due to mobile terminal movement as well as the surrounding traffic conditions. Constantly recalculating and updating the resource assignments to adapt to these changes will create a large computational load in an RAA. In addition, the measurements taken to reflect the propagation conditions may be both unreliable and costly to retrieve since they are transmitted over the air and delayed. Thus the measurement collection process is stealing some of the capacity available for user data transmission.

In most practical RAAs, the complexity is reduced by limiting the input data set. This can be done by using either partial or aggregate information. It can also be achieved by limiting the rate at which the measurements are taken. One example would be to use measured SIR values to predict future SIRs that will arise as a result of RAA actions. The rate at which measurements are taken may vary considerably. On the one hand, we can think of very fast or (near) instantaneous measurements, or, on the other hand, of very slow measurements that more reflect average propagation and traffic conditions. We can imagine the following two extremes.

Static resource allocation

Here measurements are taken at a very slow rate, or only before the system is deployed. The resource allocation is therefore based on *a priori* statistical information. This could be average propagation conditions in certain areas, e.g. around a certain access port. Further, we could estimate the average traffic load conditions and some given fixed QoS requirement. One can say that the resource allocation here is mainly done during the planning phase of system deployment.

Dynamic resource allocation

In dynamic resource allocation, we use very frequent measurements and we thereby attempt to accurately track changes in propagation conditions and user traffic requirements. A hypothetical, ideal system would operate on instantaneous (true) values, which of course would correspond to infinite measurement updates. The frequency of measurement and the rate of subsequent reassignments are dependent on the rate of change in these latter conditions. In more realistic systems with fast-moving terminals, 100 updates per second or more is a common requirement.

The schemes above can seen as archetypes. Neither of them exists in their purest form in practice. Certainly almost all practical static allocation schemes will take at least some traffic conditions into account, e.g. which mobiles are currently active. On the other hand, dynamic allocation schemes will also rely on *a priori* information to avoid excessive signaling load, e.g. large-scale propagation parameters and average path losses that are mainly dependent on the locations of the access points, which remain static. Dynamic resource allocation schemes are discussed in more detail in Chapters 4 and 7.

Exercises

2.1 Using the model in Section 2.1, let us assume that each terminal is assigned only one access port and that each waveform can be used only once by each access port. Let every user be assigned only a single waveform. Finally, when all access ports and/or waveforms are occupied, the remaining users are not assigned any waveforms or ports. What is the number of possible assignments?

2.2 In a wireless communication system, active terminal locations are distributed according to a 2D Poisson process with intensity parameter ω'. Let us assume that the terminals are independent of each other and each of them is active with probability q.

a) What will be the probability of finding exactly k terminals in the service region?
b) What is the expected number of active terminals in an area of size A?
c) Assume that five terminals are observed in some area of size A. What will now be the probability that more than three of these terminals are active?

2.3 In a wireless system for mobile telephony, active terminals making calls in a certain area are assigned one access port. Each of the access ports have in turn been assigned N channels. Calls are assumed to have exponentially distributed duration (holding times) and Poisson-distributed arrivals. A call is blocked if all N channels are busy. Show that if N is large, the number of calls in progress in the area in a certain snapshot is approximately Poisson-distributed!

2.4 In a wireless network there are two access ports and three terminals. The propagation conditions are characterized by the following gain matrix:

$$\boldsymbol{G} = \begin{pmatrix} 0.02 & 0.0005 & 0.05 \\ 0.002 & 0.01 & 0.001 \end{pmatrix} \tag{2.26}$$

Let us assume that the transmitter power is 1 unit, whereas the noise power N is 0.001 units for all terminals. Two orthogonal waveforms (channels) are available. Determine the optimal access port and channel assignments and the achieved SINRs in the various links!

2.5 Repeat Problem 2.4, but instead of two orthogonal channels, assume that infinitely many waveforms are available. These waveforms all have cross-correlations

$$\kappa_{ij} = \begin{cases} 1, & i = j \\ 0.1, & i \neq j \end{cases} \tag{2.27}$$

What are the resulting SINRs now?

References

3GPP. 1999 (December). *3GPP specification.* TS 23.107. 3rd Generation Partnership Project (3GPP).

P. Schramm, H. Andreasson, C. Edholm, N. Edvardsson, M. Hook, S. Javerbring, F. Muller and J. Skold. 1998. Radio interface performance of EDGE, a proposal for enhanced data rates in existing digital cellular systems. *Proc. 48th IEEE Vehicular Technology Conference*, Ottawa, Canada, vol. 2, 1064–1068.

3 Medium access control

3.1 Overview

In wireless networks, multiple terminals need to communicate at the same time and a medium access control (MAC) protocol allows several terminals to transmit over the wireless channel and to share its capacity. MAC protocols multiplex several data streams of different terminals to share the same channel and deal with issues such as addressing, how a terminal obtains a channel when it needs one, and so forth.

The design of MAC protocols closely relates to the condition of the physical channels. Initially MAC protocols were designed for wired communications where multiple computers need to transmit data packets at the same time in a local area network (LAN). With wired networks, the physical medium can be copper or fiber optics, which are in general very reliable with abundant bandwidth. Packet loss in wired networks is mainly due to collisions and the MAC designs are relatively simple.

The MAC design in wireless networks is much more challenging. The difficulties lie in the following aspects. With wireless communications, a radio signal may experience reflection, diffraction or scattering before reaching its receiver. Any of them will deteriorate the signal and incur variation of signal quality in time, frequency and space. Another main issue is the broadcast nature of wireless channels. For reliable transmission against fading, strong radio transmission power needs to be used by the transmitter. This incurs strong interference with other terminals in the vicinity. The stronger the transmission power or the closer the neighboring terminals, the stronger the interference will be. Because of fading and interference, wireless networks are more vulnerable compared to wired ones. Usually the bit error rate of wired networks is better than 10^{-6} and that of wireless ones is worse than 10^{-3}. The difficulty also lies in the fact that wireless terminals usually have to operate in half-duplex mode. This is because transmission power is in general much stronger than reception power, and with full-duplex operation the leakage of transmission power to the receiver component will incur very strong self-interference and therefore the terminal will not be able to receive packets or detect a collision when it is sending.

MAC schemes can be divided into two categories, contention-free and contention-based protocols. A contention-free MAC protocol requires a central controller to coordinate the resource allocation and the central controller can be a base station in a cellular network or an access point in a wireless local area network. With a contention-free MAC protocol, terminals use predetermined or assigned network resources, e.g. frequency

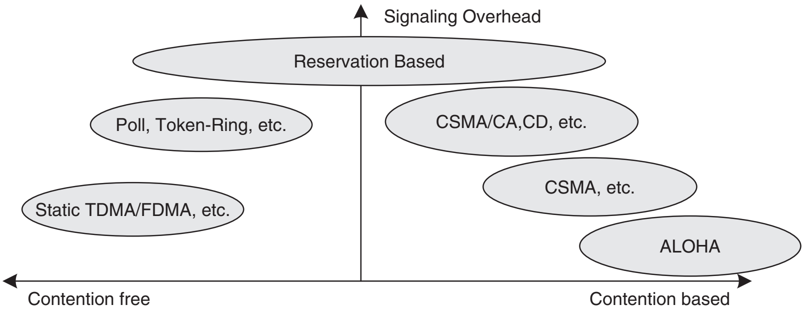


Figure 3.1 Comparison of MAC protocols.

bands, time slots, codes, antennas, and so on, to send packets. Since a central scheduler controls the channel access, the transmissions of different terminals are guaranteed to be conflict free.

Contention-based MAC protocols allow terminals to access the wireless medium randomly when they have data to send. Since there is no coordination, packet collisions are inevitable and it is essential to design the MAC protocol in such a way that collisions can be avoided as much as possible. The contention-based protocols usually perform well when the network has light traffic. With heavy traffic in the network, many terminals tend to access the channel at the same time, resulting in lots of collisions and significant performance degradation.

Contention-free MAC protocols are efficient in guaranteeing the QoS of all terminals in the network. They are also effective in improving network throughput and reducing network response time when the network is heavily loaded. On the other hand, contention-based protocols are easier to implement. They perform very well when the network has light traffic. Figure 3.1 compares all the MAC protocols that will be discussed in this chapter regarding contention level and signaling overhead. For contention-based protocols, as we move from ALOHA to CSMA, more and more signaling is introduced to coordinate the transmissions of different terminals such that collisions can be avoided as much as possible. On the other hand, for contention-free protocols, as we move from static access to reservation-based protocols, more and more signaling is introduced to enable more dynamic and flexible network resource allocation.

3.2 Data traffic and performance measures

The communication model is illustrated in Figure 3.2, assuming there are M terminals, T_0 to T_{M-1}, competing for channel access. Assume one of the terminals, e.g. T_0, is scheduled to send a source data message to the receiver, R_0. The feedback channel, z_i, is used to inform transmitter i whether or not the transmission was successful. The terminals may buffer messages while waiting for the proper wireless resources, such as time slots and frequency bands, to transmit new messages or retransmit old ones if the previous transmissions were not successful.

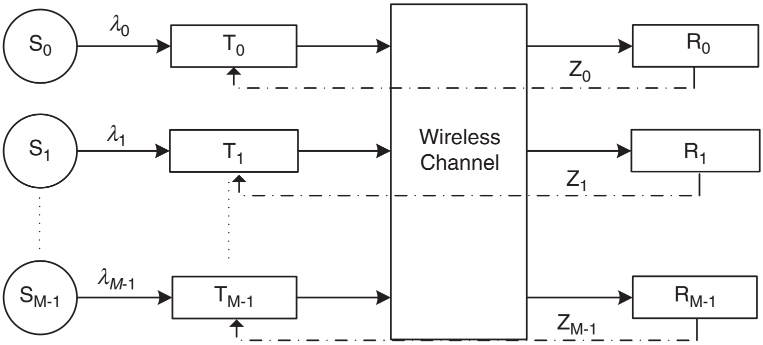


Figure 3.2 Message-oriented, multi-user communication model.

The sources, $\mathrm{S}_0, \mathrm{S}_1, \ldots, \mathrm{S}_{M-1}$, depend on application types. For example, there are almost continuous streams of packets over a long time, e.g. streaming audio, video, or file transfers; and there are applications generating sporadic data packets, e.g. web browsing or email. In many cases a bursty traffic model that generates messages at random intervals is appropriate. An example is given below.

A message may consist of a number of source symbols. Define a simple message model in which each message has N_m symbols. The messages are generated at random time instances a_1, a_2, a_3. While there are many types of stochastic point processes that can be used to model the message arrival, we mainly consider the following two arrival models:

(i) Deterministic model: The intervals between packet arrivals are the same and $t_k = kt_0$. The arrival rate is $\lambda = \frac{1}{t_0}$ messages/s.
(ii) Random model: The message arrival follows a Poisson process with rate λ messages/s.

3.2.1 Delay

We focus on three types of performance measures. First, we will use the delay. Define the arrival instant of the ith message at a terminal to be a_i and the instant when the same message is sent to the receiver b_i. The delay of the ith message is defined as

$$D_i = b_i - a_i. \tag{3.1}$$

The sequence of message delays is stochastic and the delay may be caused by a transmission delay, i.e. the time used to transmit the message with a data rate, round-trip time for receiving an acknowledgment, retransmission delay, and resource competition such as waiting time in queues for being scheduled. Assume the process is stationary and has finite moments. After sufficient time,

$$E[D_i] = E[D] = \overline{D}, \tag{3.2}$$

which is independent of i, and the system is stable. When the system is overloaded, queues may build and the delay sequence may have an average that is ever-increasing. With an infinite buffer size, the moment of D will be infinite.

3.2.2 Delivery performance

The second measure of interest is the quality of data received. There are several ways to measure the quality.

Residual error probability is the probability that a message is received with error at delivery. In most applications, this error probability has to be very low and is mostly used as a constraint. The bit error rate or bit error ratio (BER) is one measure that is frequently used. BER is the number of bits in error divided by the total number of transferred bits during a time interval. BER has no unit and is expressed as a percentage. As an example, assume an 8-bit message is sent,

$$1\ 1\ 0\ 0\ 1\ 0\ 0\ 1. \tag{3.3}$$

The received message is

$$\underline{0}\ 1\ 0\ \underline{1}\ 1\ 0\ \underline{1}\ 1. \tag{3.4}$$

The number of bit errors is 3 and the BER is $\frac{3}{8} = 37.5\%$. In practice the BER is expressed as an approximate estimate of the bit error probability, p_{BER}, which is accurate for a long run and a large number of bit errors. BER can be affected by noise, interference, synchronization problems, fading and so forth. BER can be reduced by using higher transmission power, i.e. stronger signal strength, a lower modulation order or more robust channel coding.

The packet error rate (PER) is another way. PER is the ratio, as a percentage, of the number of packets not successfully received to the number of packets sent by the transmitter. A packet is not successfully received if at least one bit in the packet is erroneous. Similar to BER, usually PER is expressed as an approximate estimate of the packet error probability, p_{PER}, which is accurate for a long run. Assuming all packets consist of the same number of bits, and in each packet bit errors are independently and identically distributed, PER can be expressed as

$$p_{PER} = 1 - (1 - p_{BER})^N, \tag{3.5}$$

where N is the number of bits in one packet.

In addition to losses at the receiver because of decoding errors, messages may also be dropped because of buffer overloading, that is, arriving messages may find the terminal buffer full and be rejected. Messages may also be dropped if they have waited too long in the buffer, i.e. the waiting time is longer than a certain threshold. Let D denote the conditional message delay, given that the message is not rejected. As shown in Figure 3.3, let $\lambda_{r,k}$ denote the dropping rate of rejected messages for terminal k. $\lambda_{r,k}$ depends on the buffer size, BER, PER, delay threshold, and so on. Some applications do not allow any message dropping at all and require the retransmission of all lost messages.

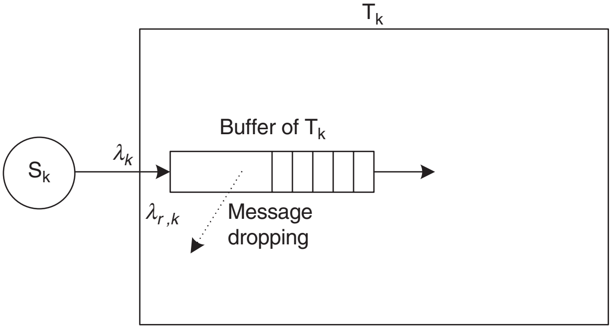


Figure 3.3 Message dropping.

3.2.3 Throughput

Another performance measure is the throughput. The individual throughput of a terminal is the expected number of messages successfully delivered to the receiver per unit time and is

$$S_k = \lambda_k - \lambda_{r,k}. \tag{3.6}$$

Throughput can be expressed in symbols/s or bits/s. A stable system with an infinite buffer size is lossless. There is no dropping of messages in this system and the throughput is equal to the arrival rate. The system throughput is defined as the sum of the individual throughputs,

$$S = \sum_k S_k. \tag{3.7}$$

The common design paradigm is to maximize the system throughput, using the user performance measures as constraints.

Alternative definitions of the individual throughput exist. One measure used for file transfers is called the link throughput and defined as

$$\tilde{\Lambda}(N) = E\left[\frac{N}{b_N - a_1}\right], \tag{3.8}$$

where N is the size of a file and the denominator is the time spent to deliver the file. In this definition, no messages are allowed to be dropped and all lost messages will be retransmitted until success. For a large N, this measure approaches the individual throughput and the link throughput can be used to estimate the individual throughput. Link throughput is used in most web browsers to indicate the transfer rate. When the message size is small or the buffering delay is large, the discrepancy between them may be large. Sometimes the normalized link delay is used, which is defined as

$$\tilde{D}(N) = \frac{1}{\tilde{\Lambda}(N)}. \tag{3.9}$$

3.3 Contention-free access protocols

Contention-free access protocols allow each terminal to send packets using predetermined resources, e.g. time slots, frequency bands or codes. Usually there is a central scheduler coordinating the transmissions of different terminals, and there will be no collisions in the network.

3.3.1 Resource assignment techniques

The resources of a wireless network can be divided into orthogonal partitions called channels and these channels can be assigned to different terminals. In the following, we introduce four basic partitioning techniques, frequency division multiple access (FDMA), time division multiple access (TDMA), orthogonal frequency division multiple access (OFDMA) and space division multiple access (SDMA). A combination of these methods can also be used to partition channels for further network performance improvement.

Frequency division multiple access

A simple resource allocation technology is fixed resource allocation, with which each link is assigned some fixed channels. The oldest way of resource allocation uses FDMA. An example is shown in Figure 3.4, where multiple terminals are communicating with the base station. In one session, a link is assigned a frequency band that is different from all other links so that the corresponding receiver can tune to the desired band to receive packets without interference. An example is audio and TV broadcast. If an FDMA channel is not in use, then it sits idle and cannot be used by other users. Every terminal will transmit simultaneously and continuously on the channel assigned. FDMA is usually implemented in narrowband systems. Its symbol time is large compared to the average delay spread. FDMA is not efficient in handling traffic flows with different bit rate requirements and it would be necessary to modify FDMA to allocate

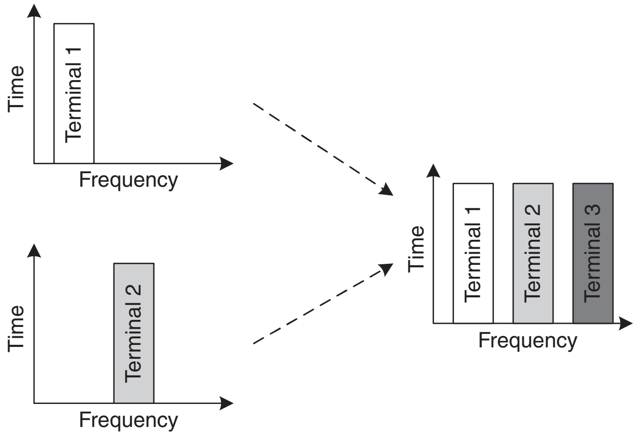


Figure 3.4 An example of fixed FDMA access.

frequency bands of different sizes to different links to accommodate the differences in rate requirements. For continuous transmission, fewer bits are needed for overhead purposes (such as synchronization and framing bits). With FDMA, both transmission and reception operate at the same time and the terminals are operated in duplex mode.

Time division multiple access

TDMA is more flexible in handling applications with different rate requirements. With static TDMA allocation, all terminals are synchronized and time is divided into non-overlapping time slots. Each link is assigned one or multiple fixed time slots such that there is only one link active at any time. For example, if there are two terminals communicating with the base station and they have the same rate requirement, one terminal can be assigned all even time slots and the other odd ones. When there are multiple terminals in the network, more advanced scheduling algorithms are needed. An example is illustrated in Figure 3.5. With TDMA, packet transmission for any terminal is not continuous and data is first buffered and then transmitted in bursts. In addition, terminals need no duplex operations since they use different time slots for transmission and reception. TDMA can allocate different numbers of time slots per frame to different terminals, allowing resources to be supplied on demand to different terminals.

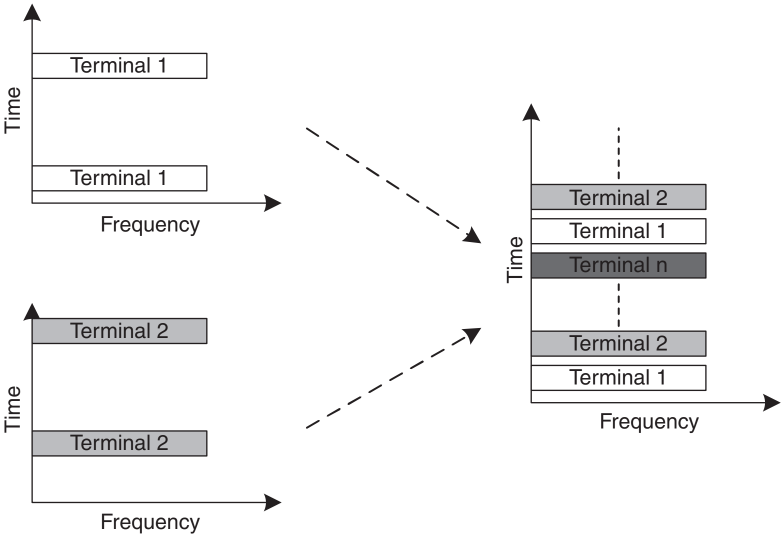


Figure 3.5 An example of fixed TDMA access.

Example 3.1: Consider a Frequency Division Duplexing (FDD) TDMA system. Each TDMA frame has a length T_c and each packet transmission takes time T. Assume the packet errors in each frame are independent. Each packet is lost with probability p. To correct errors, the stop-and-wait ARQ (Automatic Repeat reQuest) protocol is used. The downlink and uplink transmission of data and ACK/NACK (acknowledgment/negative acknowledgment) packets are illustrated in Figure 3.6 and the ACK/NACKs are always

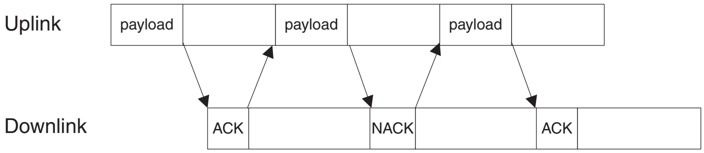


Figure 3.6 Timing in the system.

received free of errors within one TDMA frame. The uplink data transmission rate is *R Mbps*.

a) What is the expected packet service time?
b) If the transmitter always has packets for transmission, what is the expected throughput?

Solution: The probability that a packet is successfully transmitted at the *k*th attempt is

$$p_k = p^{k-1}(1-p).$$

The generating function is

$$L(z) = \sum p_k z^k = \frac{(1-p)z}{1-pz},\ (pz < 1). \tag{3.10}$$

$$L'(z) = \frac{(1-p)}{(1-pz)^2}(1-pz+pz) = \frac{1-p}{(1-pz)^2}. \tag{3.11}$$

The expected number of attempts is

$$L'(1) = \frac{1}{1-p}. \tag{3.12}$$

The expected service time is

$$\overline{D} = L'(1)T_c + T - T_c = \frac{T_c p}{1-p} + T. \tag{3.13}$$

The expected throughput is

$$\overline{S} = \frac{T}{\overline{D}}R = \frac{RT}{\frac{T_c p}{1-p} + T}. \tag{3.14}$$

Spread-spectrum multiple access

Spread-spectrum multiple access is a class of multiplexing techniques that combine both time and frequency multiplexing. These schemes are characterized by signals with a bandwidth much larger than the information data rate. The two most popular schemes in this class are termed *Frequency Hopping* (FH) systems and *Direct Sequence* (DS) systems [L. Ahlin et al., 2006]. Frequency Hopping Code Division Multiple Access (FH-CDMA) is a combined time and frequency multiple access scheme. Similar to the traditional FDMA systems, the available bandwidth is divided into a number of narrowband channels. In addition, similar to TDMA, time is also divided into time slots. The users transmit narrowband signals in one of the channels during a time slot, a *chip*. In the subsequent time slot the station keeps transmitting but on a new frequency channel. The user thus hops from frequency to frequency. Each user is assigned a unique hop sequence. The (narrowband) receiver follows the same hop sequence as the transmitter, thus tracking the transmitter in every time slot. If the hop sequences are chosen such that no chips will overlap, it is obvious that the FH signals are orthogonal and such a system will be capable of transferring the same amount of information as an FDMA or TDMA system occupying the same bandwidth. In the case when synchronization between the different users is not possible, pseudorandom hopping sequences may be preferable. These sequences can cause hits between users but can have the capability to average out interference.

Frequency hopping systems can be divided into fast or slow frequency hop. A fast frequency hop system is one in which the hopping rate is greater than the message symbol rate; in the slow frequency hop system, the hopping rate is smaller than the message symbol rate.

Direct Sequence—Code Division Multiple Access (DS-CDMA) is based on direct sequence spread-spectrum modulation. In direct sequence spread-spectrum systems (DSSS), the information signal is spread at baseband and then the spread signal is modulated by a carrier frequency in a second stage. Following this approach, the process of modulation is separate from the spreading operation. An important feature of a DSSS system is its ability to operate in the presence of strong co-channel interference. A popular definition of the processing gain (PG) of a DSSS system is the ratio of the signal bandwidth to the message bandwidth. Different users in a DS-CDMA system use different spreading waveforms, hence they may use the same carrier frequency and transmit the spread signals simultaneously. Thus there is no physical separation in time or in frequency between signals from different users. Spreading is a second modulation (after bits encoded into a digital waveform, e.g. binary phase shift keying, BPSK); direct sequence spreading codes are inherently digital. Unlike TDMA and FDMA, spread signals from different users do interfere with each other unless the transmissions from all users are perfectly synchronized and orthogonal spreading codes are employed. In general, synchronization across users is hard to achieve in the uplink of most practical wireless systems. In some situations, we may not want to restrict ourselves to orthogonal spreading codes. Therefore, the performance of DS-CDMA systems is very much dependent on MAC interference and the way it is managed.

A DSSS system can reduce the effects of interference on the transmitted information. An interfering signal may be reduced by a factor which may be as high as the processing gain. For instance, narrowband interference can be reduced by a factor up to the processing gain of the spread spectrum system. The MAC interference is reduced by the spreading code cross-correlation. That is, a DSSS transmitter can withstand more interference if the processing gain is increased.

A major disadvantage of the DSSS system is the *near–far effect*. This effect is prominent when an interfering transmitter is closer to the receiver than the intended transmitter. Although the cross-correlation between codes A and B is low, the correlation between the received signal from the interfering transmitter and code A can be higher than the correlation between the received signal from the intended transmitter and code A. So, detection of proper data becomes difficult.

Orthogonal frequency division multiple access

OFDMA is based on orthogonal frequency division multiplexing (OFDM). OFDM uses orthogonal subcarriers to send multiple data symbols in parallel and achieves high spectral efficiency. As shown in Figure 3.7, the transmitted OFDM signals, d_0, $d_1, \ldots, d_{K-1}$, can be obtained by performing the inverse fast Fourier transform (IFFT) operation on the set of data symbols, b_0, $b_1, \ldots, b_{K-1}$, to be sent on the K orthogonal subcarriers. At the receiver, the data symbols can be recovered from the orthogonal subcarriers by using the fast Fourier transform (FFT) operation.

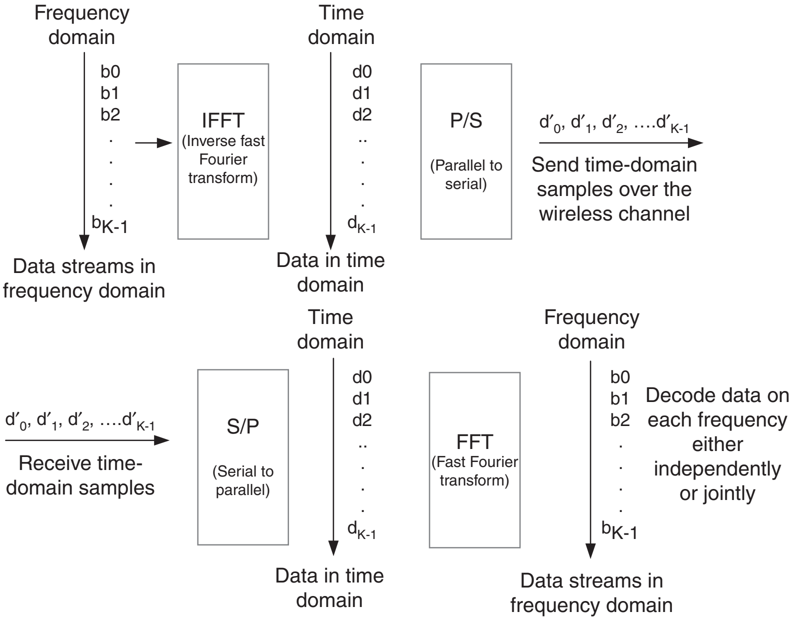

Frequency domain
b0
b1
b2
bK-1
Data streams in frequency domain
IFFT (Inverse fast Fourier transform)
Time domain
d0
d1
d2
dK-1
Data in time domain
P/S (Parallel to serial)
d′0, d′1, d′2, ...d′K-1
Send time-domain samples over the wireless channel
d′0, d′1, d′2, ...d′K-1
Receive time-domain samples
S/P (Serial to parallel)
Time domain
d0
d1
d2
dK-1
Data in time domain
FFT (Fast Fourier transform)
Frequency domain
b0
b1
b2
bK-1
Decode data on each frequency either independently or jointly
Data streams in frequency domain

Figure 3.7 OFDM modulation and demodulation using FFTs.

The spectrum of an OFDM symbol is illustrated in Figure 3.8 in which the left plot shows the waveform of one subcarrier and the right plot that of five subcarriers. As we can see in the figure, the sample subcarrier frequencies are chosen such that different subcarriers are orthogonal to each other and, while the spectra of different subcarriers overlap with each other, the cross-talk between subcarriers is eliminated. An OFDM system can therefore be considered as providing a number of orthogonal channels with channel gains h_i, $i = 0, 1, \ldots, K-1$. Adaptive modulation and coding can be employed in the frequency domain and the modulation and coding scheme (MCS) can be adjusted according to the channel gain of each subcarrier.

OFDM is very easy to implement and is robust against multipath fading and inter-symbol interference. Also, OFDM has low sensitivity to time synchronization errors. On the other hand, OFDM is very sensitive to frequency domain errors such

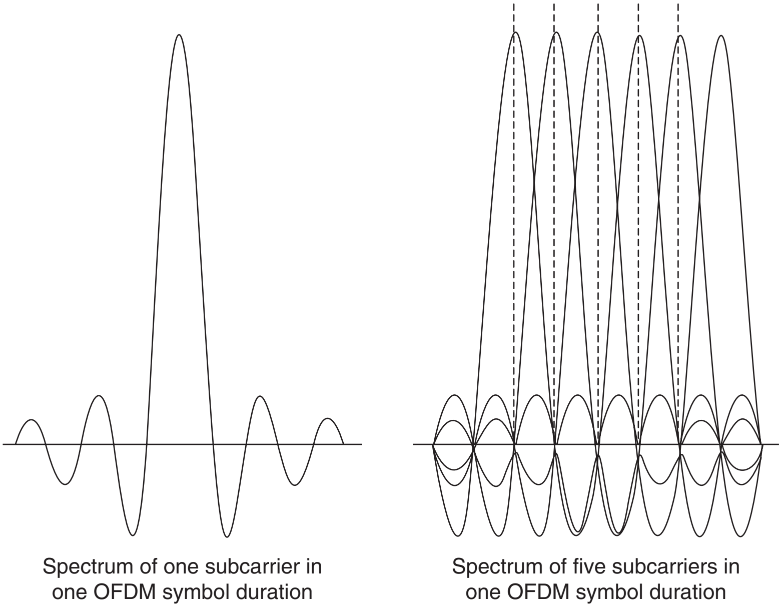

Spectrum of one subcarrier in one OFDM symbol duration
Spectrum of five subcarriers in one OFDM symbol duration

Figure 3.8 Illustration of OFDMA spectrum.

as Doppler shift and frequency synchronization problems. This is because the FFT is done at the baseband frequency, after down-converting the signals from the RF carrier frequency to the baseband using a local oscillator. Ideally the local oscillator should have the same frequency as the carrier frequency, which is determined by the transmitter. In practice, the local oscillators of the transmitter and receiver drift inevitably and frequency domain errors are hard to avoid. Another principal drawback of OFDM is a large signal peak to average power ratio (PAPR).

Multiple access is achieved in OFDMA by assigning subsets of subcarriers to individual users. For example, a subcarrier can be allocated to the user with the best relative channel condition. OFDMA therefore provides scheduling flexibility in another dimension, the frequency domain. OFDMA is a kind of FDMA in a wide sense. As shown in Figure 3.4, conventional FDMA does not allow overlapping in spectrum bands and uses guard bands to separate out adjacent bands. On the other hand, as shown in Figure 3.8, OFDMA incorporates the orthogonality without guard bands, allows overlapping adjacent spectrum bands and therefore is much more spectrum efficient. By assigning different numbers of subcarriers to different users, differentiated quality of service can be easily supported.

Space division multiple access

Mobile users are usually located far away and SDMA uses the spatial separation to reuse the frequency spectrum for higher network capacity.

The simplest form of SDMA is reusing the same frequency in different cells of a wireless cellular network. To ensure acceptable co-channel interference, the cells that reuse the same frequency should be sufficiently separated in space. This sets a limit on

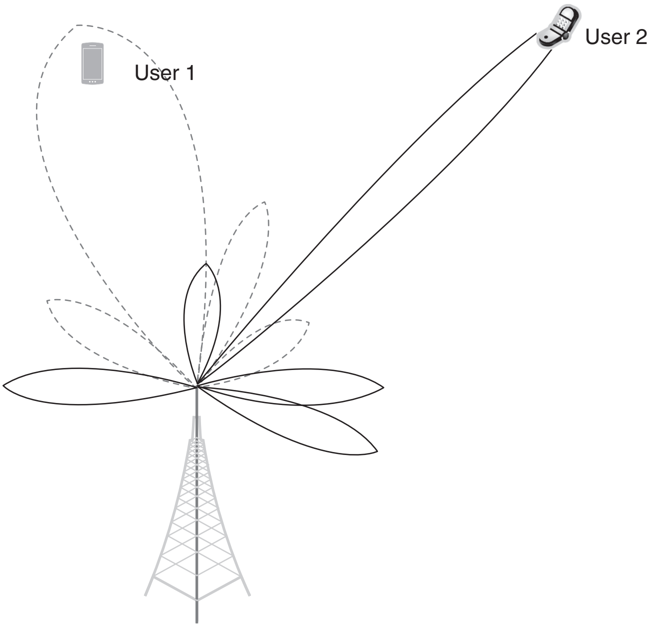


Figure 3.9 Adaptive spatial processing for two users on the same conventional channel in a cell.

how many cells a region can be divided into and the network frequency reuse factor, which will be thoroughly discussed in Chapter 5.

A more advanced SDMA technology enables frequency reuse within each cell. This technology uses smart antenna arrays and intelligent signal processing techniques to steer the antenna beam to the desired users and places nulls in the direction of other users. The frequency can be reused within each cell as long as the users are sufficiently far away. SDMA differs from FDMA, TDMA and OFDMA in that a perfect spatial separation of users cannot be guaranteed in general. The orthogonality of different users depends on the spatial correlation among the channels of the users. When spatial subchannels are not close to orthogonal, it may result in excessive co-channel interference and degraded performance. SDMA algorithms should be able to cope with this issue and multiplex data streams of different users only when their channels are sufficiently uncorrelated. The number of available spatial channels will depend on the environment and the number of transmitter and receiver antennas.

Figure 3.9 shows two users served by SDMA using the same channel in one cell. The beam pattern with the dotted lines is used to communicate with user 1 and the pattern with the solid lines with user 2. Each user is located at the actual direction of its signal. User 1 is located at a null or minimum gain point of the beam pattern of user 2 and vice versa. When the users move, the beam patterns need to be constantly updated to insure the orthogonality of the patterns.

Hybrid access

Different channel partitioning techniques can be used together to achieve finer granularity of resource management. An example is given in Figure 3.10 where both

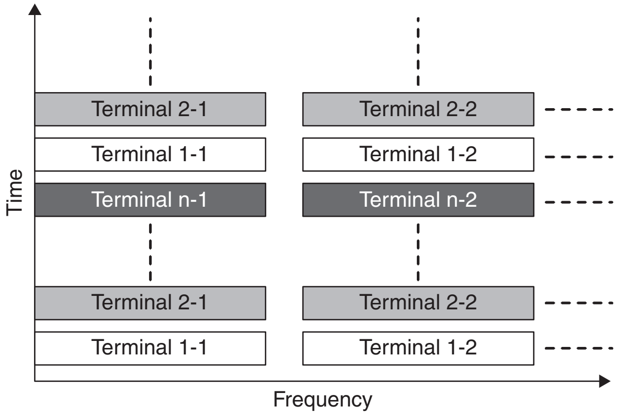


Figure 3.10 An example of hybrid access.

FDMA and TDMA are used and each terminal is assigned one resource grid in each frame.

Performance tradeoff

In the following we take a close look at the resource sharing conflicts in multi-user networks. The treatment will be slightly simplified by investigating the uplink of a single-cell system, or a multi-point-to-point system. As will be seen, the uplink problem is indeed of a different characteristic from the downlink (a point-to-multi-point system), since all resource management has to be performed by the terminals in a distributed manner. This is in particular a critical factor in systems with short messages and stringent demands on delay (response time in an interactive system) where no time or system bandwidth can be wasted for coordinating the resource management.

Consider the block diagram in Figure 3.11. The receiver is the access port trying to receive the messages from all the different terminals. To keep the models simple, assume FDMA is used. The system is lossless with infinite buffer storage in the terminals. Further assume that the symbols are transmitted at a rate $\frac{1}{T_b}$, given that the whole channel bandwidth is available. The message will thus have a duration of

$$T_m = N_m T_b \tag{3.15}$$

and the system throughput is

$$S = \frac{1}{T_m}. \tag{3.16}$$

Now explore the possible RRM strategies that may be employed. For instance, split the available bandwidth or time slots into M_0 identical, orthogonal channels using FDMA or TDMA. The symbol rate on each of these channels, $\frac{1}{T_c}$, is obviously a factor

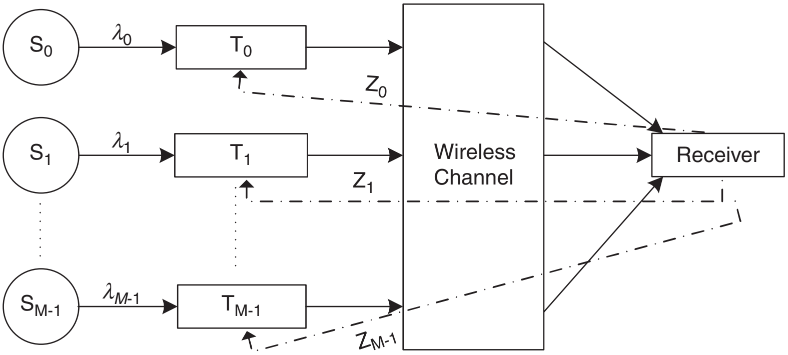


Figure 3.11 A model for an uplink multi-access system.

M_0 lower than if the entire bandwidth or time slots were used. The duration of a message in one of these channels is

$$T'_m = T_c N_m = M_0 T_b N_m = M_0 T_m \tag{3.17}$$

and the corresponding throughput is

$$S' = \frac{1}{T'_m} = \frac{S}{M_0}. \tag{3.18}$$

The messages are generated by M sources with message rates $\lambda_0, \lambda_1, \ldots, \lambda_{M-1}$. In the following, a communication system with symmetric traffic load is studied, i.e. $\lambda_i = \frac{\lambda}{M}$. The messages are fed to the M terminals, which by means of a resource sharing scheme, an access algorithm, will decide which channels to use and when. The receiver will provide feedback information, Z, concerning the outcomes of previous transmissions.

With static channel assignment, each user is tied to a channel and $M \leq M_0$. If $M = M_0$, any additional user will be rejected in the admission control as no channels are available to serve them. An admitted user may use his or her channel at any time, but will never be allowed to use the channel of other users, even if the channel is not in use. Message queues in every terminal are transmitted totally independently of the other terminals. The message transmission process of each terminal can be modeled by what is in queuing theory denoted as an $M/D/1$ queueing system. Such a system has Markovian (Poisson) arrivals, a deterministic (constant) service (transmission) time and one service unit (transmission channel). The system consists of M such independent queuing systems. This type of queuing system is well known in the literature [L. Kleinrock, 1976]. It is fairly easy to derive the expected message delay in the system. If T denotes the transmission delay in each channel, we can express the expected delay as

$$E[D_{FA}] = T + \frac{\lambda_i T^2}{2(1-\lambda_i T)}. \tag{3.19}$$

In the case with M identical stations and balanced traffic, $\lambda_i = \lambda/M$ and $T = T'_m = MT_m$. Therefore,

$$E[D_{FA}] = (M + \frac{\lambda T_m M}{2(1-\lambda T_m)})T_m. \tag{3.20}$$

3.3.2 Dynamic access protocols

The network resources can be dynamically allocated using a central scheduler to achieve better network performance. Compared with contention-based access protocols, e.g. Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA), even though there is more overhead using dynamic centralized access protocols than using CSMA/CA, performance differences are not noticeable with light traffic and are considerably better with heavy loads because contention-based protocols will spend a lot of time resolving collisions. The dynamic centralized access protocols can guarantee that every terminal will get access to the network within a given length of time. In random access methods like CSMA, terminals have to check for network activity when they want to access the network. However, in general, components of contention-free protocols are more expensive than contention-based protocols because of the complex control protocols. In addition, contention-free protocols are not very easy to extend to larger area networks.

Poll-based access protocols

Polling is an access method that designates a device as a channel access administrator, i.e. the central controller. The administrator can be a base station in cellular networks or an access point in a WLAN. The administrator queries each of the other terminals each time in some predetermined order to see whether they have information to transmit. If so, they transmit packets in the following several time slots that are predetermined.

The remaining terminals may be linked to the controller in many different configurations. One of the most common polling topologies is a star, where the controller is the hub and all the other terminals are the points of the star. To get data from a terminal, the controller sends a request for data to the terminal, and then receives the data that the terminal sends, if any. The controller then polls another terminal and receives the data from that one, and so forth. The network configuration limits how long each terminal can transmit on each poll.

Access protocols based on polling are centralized ones. The channel access time and range of data rates can be predicted and usually fixed depending on the number of terminals in the network. With the centralized control, different terminals can have different priorities and those with higher priorities can be guaranteed faster access.

The overhead of polling is high because polling uses a lot of bandwidth sending notices and acknowledgments or listening for messages. The turnaround time of polling further increases the time overhead. This overhead reduces the data rate of the channel under low loads, and its throughput.

Token-ring-based access protocols

Token-ring-based protocols also provide a distributed way to access a channel without contention.

With a token-ring-based protocol, a small frame, called the token, is passed in an orderly fashion, e.g. round-robin, from one terminal to another (see Figure 3.12). A token is a special authorizing message that temporarily gives control of the channel

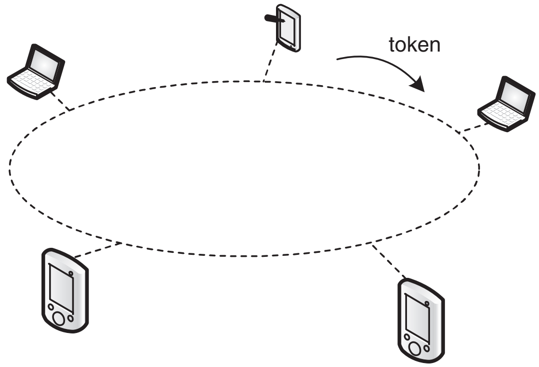


Figure 3.12 A wireless token ring network.

to the terminal holding the token. Passing the token around distributes access control among the terminals in the network. Each terminal knows from which device it receives the token and to which device it passes the token. System rules limit how long each terminal can control the token.

Terminals in the network periodically invite other terminals to join the ring by broadcasting the available resources left in the medium and the other terminals that want to will contend to join the ring.

When a terminal, B, wants to leave, it requests an adjacent terminal, A, to connect to its successor terminal, C. If A does not have a connection with C, then it connects to the next terminal in terms of the transmission order of the ring.

Performance comparison with static access protocols

We continue the discussion of the example in Section 3.3.1. Assume dynamic allocation and that in most cases there are more users than channels in the system and $M_0 < M$. An arriving message may find all channels in use. The channels are assigned to active users with messages to send and the available capacity is shared among the users demanding it. Even if $M \leq M_0$, this could happen because one of the message sources could momentarily generate more messages than the channel would be able to handle. A message arriving under such circumstances cannot be transmitted immediately and will be buffered for future transmission opportunities if some delay is allowed. In other applications when delay may be critical, e.g. telephone calls, the excess messages may be discarded. The resources and the available channel bandwidth have to be shared between the arriving and waiting messages. However, there might be problems applying this technique. Radio systems are distributed and users cannot know which other users may have messages to transmit. Usually, a part of the system bandwidth has to be assigned for exchanging this information. This assignment overhead is necessary to facilitate orderly access to the channels.

The performance of some typical FA and DA schemes is compared in Figure 3.13, which shows the normalized expected delay $E[D]$ as a function of the total arrival rate λ. At low arrival rates, only a few messages arrive from time to time. With fixed channel assignment, each user is allowed to use only its dedicated fraction of the bandwidth

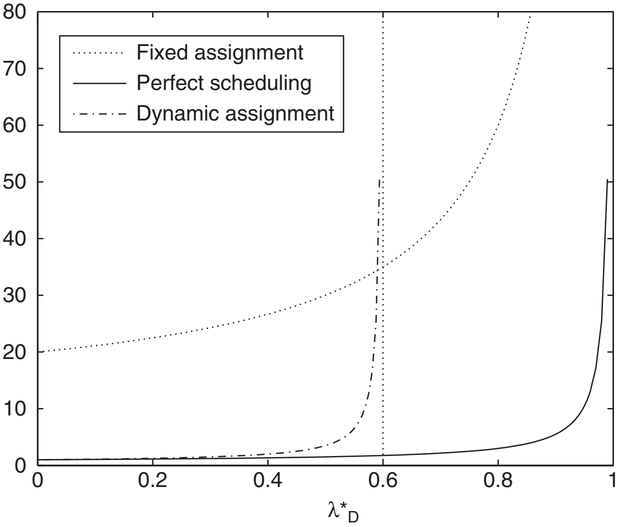


Figure 3.13 Normalized expected delay as functions of message arrival rate (M=20).

and the transmission time will be long. So fixed assignment has a comparatively high delay. At high arrival rates, the entire bandwidth is effectively utilized with fixed assignment and the maximum throughput is high. With dynamic allocation, almost all the bandwidth can be used by each user who has a message to send and the delay is very low at low traffic loads. An arriving message can be sent out almost immediately at a much higher rate than in the fixed allocation system. When the traffic load is high, more bandwidth to exchange information for resource sharing is required and the maximum throughput is much lower than in the fixed assignment system. We can see that the dynamic channel assignment algorithm achieves a low delay at moderate loads at the expense of a lower maximum throughput.

The amount of assignment overhead depends on how rapidly traffic demands change. A system with short bursty messages has a much larger overhead than a system with long messages, because the allocation is repeated much more frequently. In Figure 3.13, the assignment overhead is about 50% of the overall system bandwidth. Figure 3.13 also illustrates the performance of a Perfect Scheduling (PS) scheme, a hypothetical system with no assignment overhead. This system assumes all users could instantaneously know the buffer statuses of all other users and the channel allocation results. So messages could be transmitted without loss or overhead. The performance of this system serves as a lower bound on the expected message delay. The messages of all users in the PS system are in effect placed in a single queue. According to (3.19), and with $T = T_m$ and $\lambda_i = \lambda$, we have

$$E[D_{PS}] = \left(1 + \frac{\lambda T_m}{2(1 - \lambda T_m)}\right) T_m. \tag{3.21}$$

Therefore,

$$E[D_{PS}] = \frac{E[D_{FA}]}{M}. \tag{3.22}$$

We can see from Figure 3.13 that the asymptote of the delay, the maximum throughput, is however the same as that of the fixed assignment scheme.

Almost all communication systems nowadays implement dynamic assignment schemes and adapt the resource allocation based on instantaneous demand. These schemes usually belong to the reservation type of scheme. With a reservation scheme, part of the bandwidth is used to exchange information on traffic demands, channel states and so on, and the rest for data transmissions. Reservation-based schemes work well with long messages as the assignment overhead is negligible. With short messages, significantly more bandwidth is used for reservation and the throughput would be very low.

3.4 Contention-based access protocols

In the following, we study the distributed schemes, called contention-based access protocols, under which every user decides its channel access based only on its own observations. However, when several users decide to access a channel simultaneously, conflicts occur and none of the messages are regarded as received correctly. Compared to collisions, noise will be rather harmless and initially we will neglect the impact of noise. After a transmission attempt, each user will receive the outcome by using an almost error-free feedback mechanism, e.g. the transmission of acknowledgment messages. The acknowledgment messages are usually very short and the corresponding bandwidth can be neglected. The feedback is crucial in channel access and the feedback signal is indeed the only information available for users to make transmission decisions.

Contention-based access protocols allow terminals to access channels randomly when they have packets to send. With random access or contention-based methods, no terminal is superior to another station and none is assigned the control over another. No terminal permits, or does not permit, another one to send. At each instance, a terminal that has data to send uses a procedure defined by the protocol to make a decision on whether or not to send.

We will first introduce the ALOHA protocol, which is a pioneering MAC protocol in computer networks. The first version of the ALOHA protocol is pure ALOHA, under which each terminal sends the packet whenever it wants. Slotted ALOHA improves on pure ALOHA by dividing time into slots, and transmission occurs only when a new slot starts. With this improvement, the chance that different transmissions collide is reduced and the network throughput is improved. In this chapter, we will also introduce more complicated and advanced contention-based MAC protocols, e.g. Carrier Sense Multiple Access (CSMA), CSMA with Collision Avoidance, and IEEE 802.11 MAC, opportunistic random access. We will see how MAC evolves to handle collisions for better network performance.

3.4.1 ALOHA

In the late 1960s, the ALOHA protocol was first developed at the University of Hawaii and used in a VHF radio system. It connected terminals on the remote islands with a central computer site. Each terminal was equipped with a radio interface, as were the main computers. Thus, all the terminals belonging to a main computer were connected to the computer via a shared medium, the air interface. In the original protocol, now called pure ALOHA, the stations were not synchronized. The control signaling, i.e. ACK (acknowledgment), is sent on an independent control channel. Pure ALOHA is quite simple. For each terminal in the network, it runs the following algorithm:

- Step 1: If there is a message to send, send it;
- Step 2: If the transmission succeeds, remove the message from the queue and go to Step 1. If the transmission fails, wait a random time interval, i.e. backoff randomly, and go to Step 1.

Transmission failures are mainly due to collisions. It is crucial that the terminals do not wait for identical time intervals if a collision happens, as otherwise this would almost certainly cause another collision. To avoid this, all terminals will perform a backoff scheme and decide randomly when to transmit next time. The quality of the backoff scheme determines the efficiency of pure ALOHA and network capacity. When the total traffic is low, pure ALOHA works well; otherwise, the performance is poor.

Assume all frames have the same length, 1. Suppose the ith frame is sent starting from t_0, as shown in Figure 3.14. In order to have a successful transmission, no other terminals should transmit any signal between t_0 and t_0+1. In addition, no other terminal should send a frame between t_0-1 and t_0 because otherwise, the latter part of this frame will overlap with the ith frame. So two consecutive frame durations are needed for sending one frame successfully in a network.

A simple improvement of the basic ALOHA protocol is slotted ALOHA. With slotted ALOHA, all terminals are synchronized and time is assumed to be slotted. Terminals can send packets only at the beginning of slots with durations equal to or larger than the packet transmission time, as shown in Figure 3.15, and therefore collisions are reduced. We can see that with slotted ALOHA, collisions happen only if two terminals send

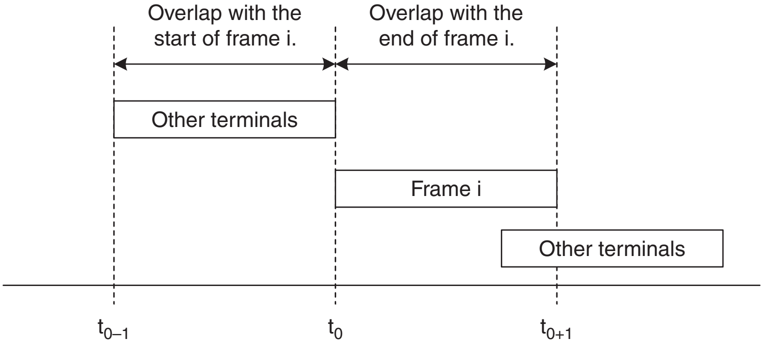

Overlap with the start of frame i.
Overlap with the end of frame i.
Other terminals
Frame i
Other terminals
t_{0-1}
t_0
t_{0+1}

Figure 3.14 Collision in pure ALOHA.

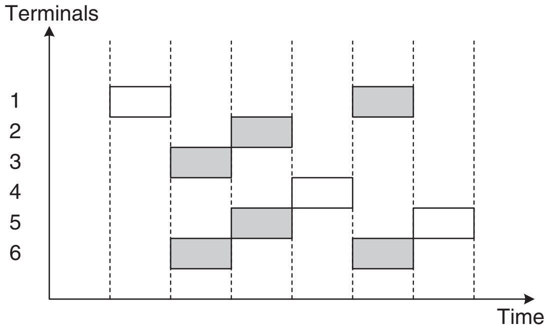


Figure 3.15 Slotted ALOHA.

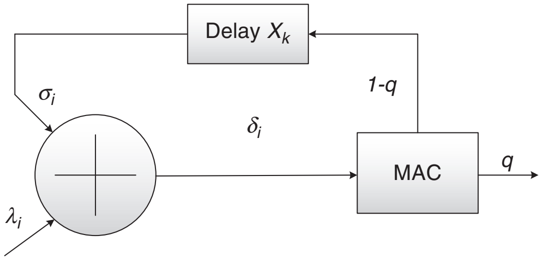


Figure 3.16 Simplified transmitter model in a slotted ALOHA system.

frames in the same slot. So only one frame duration is needed for sending one frame successfully. The efficiency of slotted ALOHA is therefore twice that of pure ALOHA.

Let us analyze the performance of a slotted ALOHA system. For this system, suppose each terminal has a packet to transmit with probability p. An exact analysis is beyond the scope of this book and some further simplifications and approximations will be made. The results will still describe the system performance fairly well. Figure 3.16 describes the message flows in one of the many (M) terminals in the network. The message arrivals follow a Poisson process with rate $\lambda_i = \lambda/M$. After arrival, a message is transmitted in the next time slot and so the transmission probability p is determined by the arrival process. Depending on the transmission attempts of the other users, the transmission may or may not succeed. If it succeeds, the message will leave the system. A message that fails transmission is stored in the buffer. After a random number of X_k time slots, message k will be retransmitted. X_k is assumed to be independent and identically distributed. Further assume that the messages that are retransmitted also follow a Poisson process with rate σ_i. This process is independent of the arrival process. With this assumption, the message arrivals are memoryless and independent of whether a message is new or retransmitted. This approximation is accurate if

$$E[X_k] = E[X] = \xi \gg 1, \tag{3.23}$$

where X denotes the stationary process X_k. The compound process of new and retransmitted messages also forms a Poisson process with intensity δ_i,

$$\delta_i = \lambda_i + \sigma_i. \tag{3.24}$$

δ_i determines the actual tries of packet transmissions and is called the access attempt rate.

Let q denote the probability of a successful transmission in a time slot for terminal i. In a stable equilibrium, the average number of messages arriving in unit of time has to be equal to the average number of departing messages, otherwise the buffer will be either empty or ever increasing in the number of packets stored. Then we have

$$q\delta_i = \lambda_i. \tag{3.25}$$

The condition for a transmission to succeed is that no other terminals are transmitting at the same time. As all terminals are identical, every one has the same transmission probability, denoted by p. Thus we have

$$q = (1-p)^{M-1} \approx (1-p)^M. \tag{3.26}$$

The probability p is given by

$$\begin{aligned} p &= \Pr\{\text{at least one arrival in the slot}\} \\ &= 1 - \Pr\{\text{no arrival in the slot}\} \\ &= 1 - e^{-\delta_i}. \end{aligned} \tag{3.27}$$

Together with (3.26), we have

$$q \approx (1-p)^M = e^{-M\delta_i} = e^{-\delta}, \tag{3.28}$$

where $\delta = M\delta_i$. Combining (3.25) and (3.28) yields

$$\lambda_i = q\delta_i = \delta_i e^{-\delta}. \tag{3.29}$$

The overall network throughput is

$$\lambda = \sum_{i=0}^{M} \lambda_i = M\delta_i e^{-\delta} = \delta e^{-\delta}. \tag{3.30}$$

This characterizes the relation between the total rate of traffic arrivals, λ, and the total rate of transmission attempts, δ, and is illustrated in Figure 3.17. The right-hand side of (3.30) is upper bounded and the maximum of λ is given by

$$\lambda \le \max \delta e^{-\delta} = e^{-1}. \tag{3.31}$$

So the maximal throughput is $\lambda^* = e^{-1} \approx 36.8\%$.

The message delay may also be estimated. Let N denote the number of retransmission attempts required for a successful transmission. A retransmission attempt consists of one time slot for the packet transmission and a stochastic delay X. The normalized delay is

$$\begin{aligned} E[D] &= \sum_{n=0}^{\infty} E[D|N=n]p[N=n] \\ &= \sum_{n=0}^{\infty}(1+nE[X])p[N=n] \\ &= \sum_{n=0}^{\infty} p[N=n] + E[X]\sum_{n=0}^{\infty} np[N=n] \\ &= 1+\xi E[N]. \end{aligned} \tag{3.32}$$

The success probability of a transmission is q and N is geometrically distributed. Assume successive transmission attempts are independent, which is reasonable when $\xi \gg 1$. The expectation is

$$E[N] = \frac{1}{q} - 1. \tag{3.33}$$

Therefore we have

$$E[D] = 1+\xi\left(\frac{1}{q}-1\right) = 1+\xi(e^{\delta}-1). \tag{3.34}$$

Together with (3.30), we have obtained the delay as a function of the arrival rate; with an intermediate parameter δ. Solving for δ in (3.30) results in not only one but two solutions, which are illustrated in Figure 3.17. Let $\delta_1(\lambda)$ and $\delta_2(\lambda)$ denote these two solutions. Then we have two possible delays, $D_1(\lambda)$ and $D_2(\lambda)$. When the system

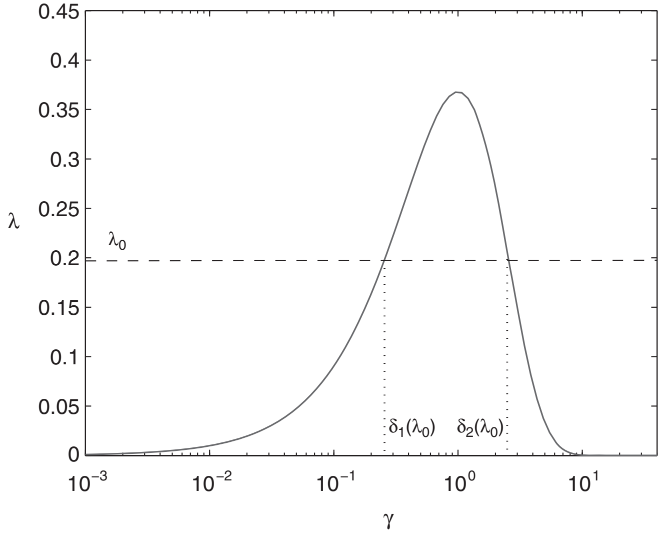


Figure 3.17 Slotted ALOHA.

achieves the maximum throughput $\lambda^* = e^{-1}$, the corresponding delay is

$$E[D] = 1 + \xi e \approx \xi e. \tag{3.35}$$

There are two states of equilibrium in general. Correspondingly there are two possible delays for every message arrival rate; one is comparatively low and the other is considerably higher in general. The equilibrium corresponding to $D_1(\lambda)$ has a low channel activity with most transmissions being successful. In the other one, there are many backlogged messages in the buffer and frequent retransmissions happen. While the attempt rate δ is high, the success probability q is low. Because of the stochastic variations in the traffic arrivals, the system moves between these two equilibrium states. It can be shown that the probability of staying in the high delay state decreases with higher ξ. So a large ξ is beneficial in stabilizing the system, but will generate a large average delay, as shown in (3.34). A more comprehensive analysis of the stability properties of ALOHA systems can be found in [R. Rom and M. Sidi, 1990].

Example 3.2: In a taxi management system, a number of taxis send requests to a central site. Each request message has 300 bytes (2400 bits). There are 60 taxis, each making requests following a Poisson process with a rate of one request per minute. The maximal data rate is 14,400 bps (bits per second). Either a fixed assignment scheme using 60 slots per frame or a random access system using a slotted ALOHA algorithm can be implemented. Assume the average retransmission delay in the slotted ALOHA algorithm is ξ = 25 and the system is in the low delay equilibrium.

a) Which system achieves the lowest expected message delay?

b) At what rate should the taxis send their requests so that the systems have the same average message delay?

Solution: The message duration and the normalized arrival rate are

$$T_m = NT_b = 2400/14400 = 1/6s$$

and

$$\lambda = 60 * 1/60 * 1/6 = 1/6.$$

a) ALOHA: Solve $\lambda = \delta e^{-\delta}$ $\lambda = 1/6$ for δ

$$\delta \approx 0.204.$$

$$E[D_A] = 1 + \xi(e^{\delta} - 1) \approx 6.7 \text{ slots} \approx 1 \text{ sec.}$$

TDMA: According to (3.20),

$$E[D_T] = MT_m(1 + \frac{\lambda T_m}{2(1-\lambda T_m)}) \approx 66T_m \approx 10 \text{ sec.}$$

So the ALOHA system has the lowest average delay.

b)

$$D_A(\lambda_0) = D_T(\lambda_0).$$

So

$$\lambda_0 \approx e^{-1} \approx 0.367 \approx 2.2 \text{ requests}/(\text{minute} \cdot \text{mobile}).$$

When the ALOHA system is used, the taxi management system is close to becoming unstable. On the other hand, it is always stable using the fixed assignment scheme for all $\lambda < 1$, i.e. for arrival rates less than 6 requests/(minute · taxi).

3.4.2 Carrier sense multiple access

To further improve network performance, we can abort transmissions as soon as a collision is detected. This is because once two packets start colliding, it is useless to continue their following transmissions as their respective receivers will not be able to decode the packets and will lead to a waste of bandwidth and energy. Hence, it is necessary to detect collision while transmitting a packet. This motivates the design of Carrier Sense Multiple Access (CSMA).

CSMA is popular in asynchronous (non-slotted) wireless networks with low propagation delays. Each terminal is equipped with a receiver to monitor if any transmission is in progress in the channel. The terminals simply measure the signal level to detect transmission, that is, carrier sensing. If there is a carrier present, i.e. an existing transmission in progress, the station defers its transmission attempt to a randomly determined later time. Otherwise, it transmits its packet. Once the packet is transmitted, the terminal waits for an acknowledgement. If no acknowledgement is received later, the packet will be scheduled for future transmissions. Like ALOHA, CSMA requires an independent control channel for acknowledgment transmissions.

There are three types of CSMA algorithms, non-persistent, 1-persistent, and *p*-persistent.

Non-persistent CSMA

Each terminal runs the following protocol:

- Step 1: If the channel is sensed idle, transmit a packet immediately;
- Step 2: If the channel is sensed busy, wait a random amount of time and go to Step 1.

The random backoff in Step 2 reduces the probability of collisions. On the other hand, if the random backoff time is too long, the channel capacity may be wasted as no other terminals may transmit either.

1-persistent CSMA

Each terminal runs the following protocol:

- Step 1: If the channel is sensed idle, transmit a packet immediately;
- Step 2: If the channel is sensed busy, continue sensing until the channel is idle; then transmit a packet immediately.

The protocol is called 1-persistent because a terminal will transmit with probability 1 whenever it finds the channel idle. With 1-persistent CSMA, if two or more terminals want to send packets while the channel is busy, their transmissions will always collide as they all begin transmission simultaneously as soon as the channel becomes idle.

p-persistent CSMA

Each terminal runs the following protocol:

- Step 1: If the channel is sensed idle, transmit a packet with probability p; with probability $1-p$, delay its transmission by one time slot and repeat this step.
- Step 2: If the channel is sensed busy, continue sensing until the channel is idle; then go to Step 1.

1-persistent CSMA is a specific case of p-persistent CSMA. The p-persistent protocol improves the 1-persistent protocol by reducing the likelihood of packet collisions through random backoff of packet transmissions.

The selection of p determines the performance of p-persistent CSMA. Suppose there are N terminals having packets to send while the channel is busy. Np is the expected number of terminals that will send packets once the channel is idle. If $Np > 1$, a collision is expected. So the choice of p must ensure that $Np < 1$ to avoid collision and $p < \frac{1}{N}$, where N is the maximum possible number of active terminals at the same time.

Effect of detection delay

An important parameter in a CSMA system is the detection delay. As shown in Figure 3.18, this is the time span between when a terminal decides to start a transmission at time t until the instant $t + t_0$ when all terminals become aware of this transmission (and thus avoid a collision). This delay includes the carrier measurement delay, the switch time from reception to transmission, and the propagation delay. If the detection delay is large compared to the message size, the carrier sensing information will be of little use since it is only able to describe the state of the channel some time ago. The capability to prevent collisions deteriorates and two terminals may both sense that the channel is empty and start their respective transmissions. However, if the detection delay approaches zero, almost all collisions can be avoided.

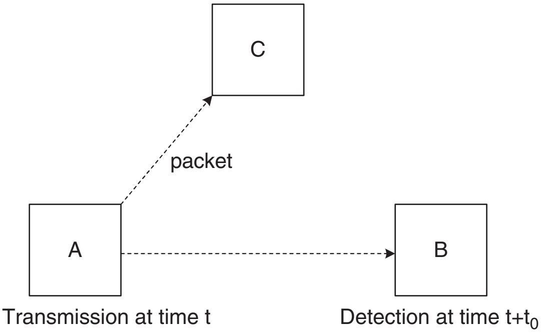


Figure 3.18 Delay in channel sensing.

3.4.3 CSMA with collision detection

With CSMA, if a terminal decides to send a packet at a certain time t, it will send the whole packet. If there is a collision with other terminals, the resources, time, frequency and energy used in sending the whole packet are wasted. CSMA/CD is designed to reduce such waste.

With CSMA/CD, each transmitting terminal will also monitor its own transmission. If it detects a collision, it stops transmission immediately and instead sends a jam signal to indicate that there has been a collision. The channel is therefore much more efficiently used since the bandwidth transmitting the entire frame is not wasted. The detection techniques depend on the channel media. For example, with electrical wires, collision detections are done by comparing the transmitted packet with the received one.

In addition to the collision detection capability, the backoff window also grows to reduce collision probabilities. So after a collision, a terminal involved in the collision will retransmit its packet after waiting for a random time period. If another collision occurs, the time window from which the random waiting time is selected is increased step by step. One popular way of increasing the backoff window is called exponential backoff. In one example, in the first collision, suppose the backoff time is chosen randomly from $[0,t]$. In the Nth collision, the backoff time will be chosen randomly between $[0,t_N]$, where $t_N = \min(2^{N-1}t, t_{Max})$, where t_{Max} is the maximum backoff window size.

CSMA/CD is easy to implement and is widely used in wired networks.

3.4.4 Carrier sense multiple access with collision avoidance

CSMA/CD is not appropriate for wireless networks. CSMA/CD detects collisions at senders, not receivers. In wireless networks, a signal should reach the receiver without any collision for successful transmission. So the receiver needs to detect collisions, which is different from CSMA/CD. Further, the signal strength is almost the same for wired networks but varies dramatically in wireless networks depending on the wireless channel between the transmitter and the receiver. A collision at the receiver in many cases will not be detected by the sender. More importantly, in wireless networks, the transmission power is usually much higher than the reception power and collision detection by the sender is not possible in practice. Some more issues are described below.

Hidden and exposed terminal problems

The hidden terminal problem is illustrated in Figure 3.19, where the arrows stand for the traffic flows and the dashed circles the communication ranges of A and B respectively. A and B are sending packets to C but A and B are out of the transmission range of each other. They will both sense the channel idle and send packets to C. Their packets collide at C. So A and B are hidden terminals to each other.

The exposed terminal problem is illustrated in Figure 3.20, where the dashed circles stand for the communication ranges of A and C respectively. Obviously A and C can

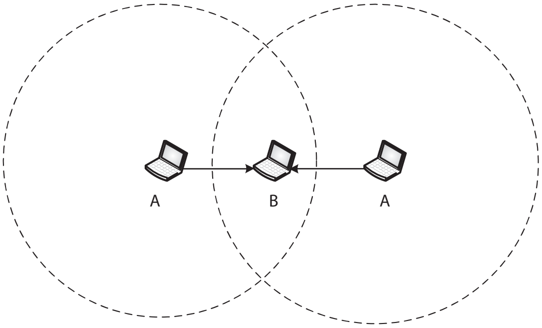


Figure 3.19 The hidden terminal problem.

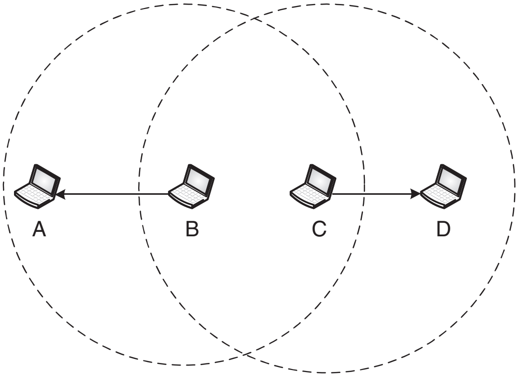


Figure 3.20 The exposed terminal problem.

send packets to B and D respectively at the same time without any collision at the receiver sides. However, if A sends a packet to B, C will sense the channel busy and will not transmit. Half the channel capacity is therefore wasted.

Carrier sense multiple access with collision avoidance

CSMA/CA differs from CSMA/CD in that collisions are avoided. A terminal with a message to transmit will sense the channel first. If the channel is idle for a short period called the DIFS (DCF inter-frame space, where DCF is the distributed coordinated function) period, the terminal can transmit. Otherwise, the terminal will defer and continue monitoring the channel until the channel is idle for the DIFS period. Then a random backoff counter within the contention window will be generated before actually sending the message. The backoff counter counts down as long as the channel remains idle. The counter will pause if the terminal detects the channel busy and resume when it detects the channel idle for another DIFS period. The terminal will transmit immediately when its backoff counter is zero. CSMA/CA can use a binary exponential backoff algorithm to control the contention window size of each terminal and resolve collisions. Each terminal has a contention window size w that has a minimum value W_{min} and a maximum value W_{max}. Before a new transmission, the backoff counter is chosen

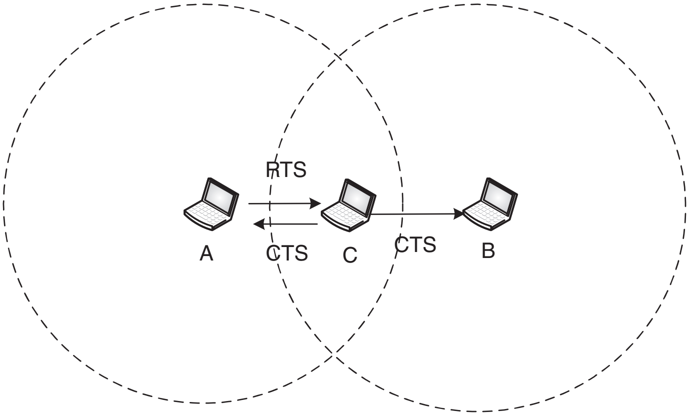


Figure 3.21 RTS/CTS exchange.

uniformly from $[0, w-1]$. At the first transmission attempt, $w = W_{min}$. After each failure, w is doubled until it reaches W_{max}.

Optionally, but almost always implemented, CSMA/CA can use a Request To Send / Clear To Send (RTS/CTS) exchange mechanism to avoid collisions. An example is illustrated in Figure 3.21. When a terminal, A, intends to send a message to C, it will sense whether the channel is idle and back off if the channel is busy. If the channel is idle, it will send a small RTS message to the intended receiver, C. If the receiver senses the channel is clear and receives the RTS, it immediately replies with a small CTS message to A. If A receives the CTS, it sends the data to the receiver. Both the RTS and CTS messages carry the time length to transmit the data message. After hearing the CTS, B will no longer transmit anything when A sends the package and therefore a collision is avoided. If A does not receive a CTS message, it starts the RTS procedure again.

RTS/CTS also solves the exposed terminal problem. Consider the traffic flows in Figure 3.20. Terminal A first sends to B an RTS, which will also be heard by C. C will therefore defer its RTS transmission to D. However, later C will not receive a CTS from B and knows that B is not within its communication range. Then C can start its transmission to D.

Example 3.3: One way to solve the hidden terminal problem is to let the receiver send a busy signal on an independent control channel. The time is slotted. When the transmitter wants to send a packet to the receiver, the transmitter checks if the control channel is idle for one slot, defined as slot t. If so, in the next time slot, $t+1$, it sends a request on the control channel to the receiver. After receiving the request, the receiver sends the busy signal on the control channel continuously until the transmitter finishes its packet transmission. After receiving the busy signal, the transmitter starts sending data at time $t+2$ on the data channel. All other transmitters hearing the busy signal keep silent. Assume two transmitters are sending data to the same access point. In each idle slot, each transmitter would like to send a packet with probability p. Each packet transmission lasts M slots. What is the throughput of the network?

Solution: For each transmitter, the throughput is $S = \frac{E}{B+I}$, where E is the expected time for successful packet transmissions, I the expected idle time, and B the expected busy time, including both collisions and successful transmissions.

$$\begin{aligned}\Pr(I = k) =& \Pr(\textit{no arrivals in the first } k-1 \textit{ slots}) \times \\ & \Pr(\textit{at least one arrival in the last slot}) \\ =& [(1-p)^2]^{k-1}(1-(1-p)^2).\end{aligned} \tag{3.36}$$

The expected idle time is

$$I = \sum_k k[(1-p)^2]^{k-1}(1-(1-p)^2) = \frac{1}{1-(1-p)^2}. \tag{3.37}$$

$$\begin{aligned}& \Pr(\textit{k busy transmissions}) \\ =& \Pr(\textit{At least one arrival in the first} \\ & k-1 \textit{ slots}) \times \Pr(\textit{no arrival in the last slot}) \\ =& [1-(1-p)^2]^{k-1}(1-p)^2.\end{aligned} \tag{3.38}$$

The expected number of busy transmissions is

$$B_N = \sum_k k[1-(1-p)^2]^{k-1}(1-p)^2 = \frac{1}{(1-p)^2}. \tag{3.39}$$

Each failed transmission takes two slots and each successful one takes $M+2$ slots. Therefore, the expected busy time is

$$B = B_N(2*(1-p_s) + p_s(M+2)), \tag{3.40}$$

where the success probability is

$$\begin{aligned}p_s =& \Pr(\textit{only one transmitter has a packet} | \\ & \textit{at least one transmitter has packet arrivals}) \\ =& \frac{2p(1-p)}{1-(1-p)^2}.\end{aligned} \tag{3.41}$$

The expected time for successful packet transmissions is

$$E = B_N p_s M = \frac{2p}{(1-p)(1-(1-p)^2)} M. \tag{3.42}$$

The network throughput follows immediately.

Conflict resolution algorithms

The algorithms introduced so far resolve conflicts by postponing transmission for a random amount of time while hoping for the best. Systematic schemes can be designed to resolve conflicts. Some examples are the tree-search algorithms proposed in [J. J. Capertanakis, 1979] and the channel-aware algorithm proposed in [G. W. Miao et al., 2012], which not only resolves conflicts but also schedules the terminals with the best channel states for data transmission. In the following we introduce briefly the tree-search algorithm.

Let all users be marked with unique integer identifications (IDs) $0, 1, 2, \ldots, M < b^K$ in the base b, that is, every identity is a K multidigit number in which each digit is among $0, 1, \ldots, b-1$. There are two modes in the system, a free access mode and a conflict resolution mode. In the free access mode, the users will send packets when they arrive. Messages involved in a collision may participate in the conflict resolution mode. The conflict resolution mode is started as soon as a conflict occurs. The system returns to the free access mode only if the conflict has been resolved.

A conflict resolution algorithm (CRA) can be used in the conflict resolution mode to sort out the order of transmissions of users in conflict. After the first collision, the users with a zero as the final digit in their IDs are allowed to transmit in the next slot. If another collision occurs, users whose final digits are 00 will transmit. If there is no collision in the next slot, indicating either no or a successful transmission, users whose IDs end with 1 are allowed to transmit. Each collision increases the number of ending digits to partially solve the conflict. This algorithm is illustrated in Figure 3.22, in which each node corresponds to the transmission in a time slot. The root is the free access mode. After a collision in a time slot or a node, b sub-branches of the node are created. If there is no collision in a slot, it forms a leaf node in the tree. When all leaf nodes are reached, the conflict is resolved and the system returns to the free access mode. So the goal is to exhaustively search the tree to allow all users to finish their data transmission. Different search algorithms can be applied, e.g. depth first or width first schemes.

The maximum number of nodes to be searched, i.e. the maximum number of slots needed to completely resolve a conflict, is

$$N_{max} = \sum_{i=0}^{K} b^i = \frac{b^K - 1}{b - 1}. \tag{3.43}$$

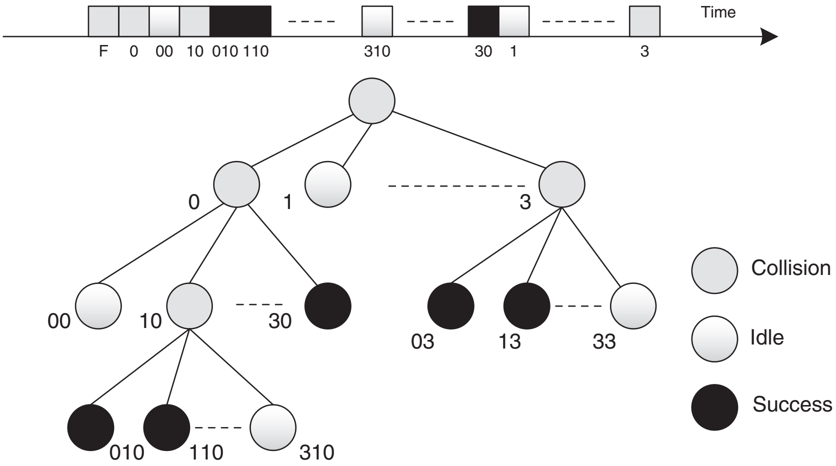


Figure 3.22 CRA with tree search.

The choice of b determines the delay and maximum throughput. With a rather large b, the free access state is skipped and CRA is almost always used. This corresponds to a multichannel system where $M_0 = b$. The users access their subchannels according to the final digit of their IDs. The same CRA is used on each subchannel if multiple users have the same final digit. b can be chosen adaptively to match the traffic load. When $b = M$, the system has a pure fixed assignment.

Reservation-based protocols

It is very costly to lose a long data packet in a collision, and reservation-based protocols can be used to avoid collisions. There are two modes, a reservation phase and a data phase, in a reservation system. A typical example is shown in Figure 3.23, where five terminals are accessing the channel. R stands for the reservation phase and the numbers stand for the data transmission phases of the corresponding terminals. The reservation phase is used to reserve the resources, e.g. time slots, frequency bands and codes, for all users to transmit in the data phase without any collisions.

There are mainly two categories of reservation schemes, scheduling and contention based. In the scheduling-based scheme, a central controller will collect necessary information from all users in the network and then apply a certain scheduling algorithm to decide the resources in the data phase allocated to each user in the network. In this case, it is not a contention-based protocol but is contention free. The scheduling-based reservation scheme is the most popular medium access scheme in cellular networks because of its high efficiency. In Chapter 4, we will discuss in more detail how scheduling-based reservation schemes are implemented in practice.

In this section, we will focus on contention-based reservation schemes, with which terminals compete for channel resources using short reservation packets, before actual data transmission. During the reservation phase, some type of contention resolution algorithm, e.g. ALOHA or CRA, is used to send reservation packets. The reservation packets should be as small as possible and carry information regarding only the terminal identity, number of resources needed, QoS requirement and so on. After the reservation phase, either a central controller announces the result or the terminals deduce the results by themselves using distributed algorithms. The result is the reservation of resources, e.g. time slots, frequency carriers and antenna ports, in the following data phase. In the data phase, messages are transmitted on the resources reserved and there are no collisions.

When slotted ALOHA is used in the reservation phase, the scheme is called reservation ALOHA. As illustrated in Figure 3.24, six terminals are accessing the channel. There are five mini-slots in the reservation phase and a corresponding five slots in the data phase. Each terminal uses slotted ALOHA to send reservation packets in a random slot out of these five mini-slots. In Figure 3.24, terminals 1, 2 and 3 send the

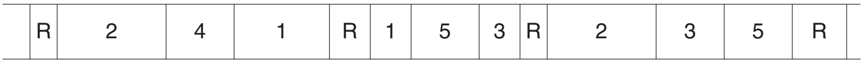


Figure 3.23 A typical slot layout of a reservation system.

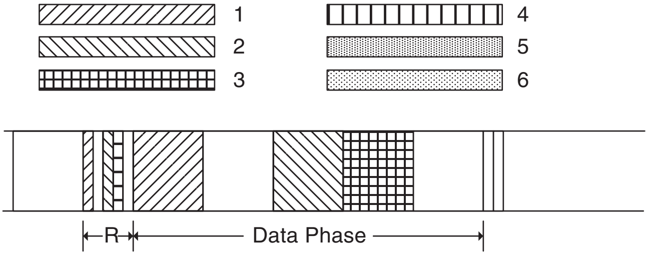


Figure 3.24 Reservation ALOHA.

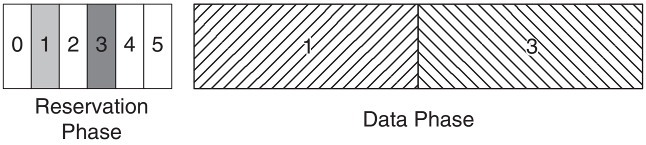


Figure 3.25 An example of the bit-map protocol.

reservation packets successfully on mini-slots 1, 3 and 4 respectively and therefore will transmit in the corresponding data slots. Both terminals 4 and 5 transmit in mini-slot 2 and a collision happens. No terminals send any reservation packet in mini-slot 5. Therefore there is no transmission in slots 2 and 5.

Another simple reservation-based protocol is the bit-map protocol. Assume there are N terminals in the network and they are numbered from 1 to N. There is a contention period of N slots. Terminal i has one corresponding slot i in the contention period. Terminal i will send a signal in the ith slot if it wants to send a frame. All terminals see all reservation signals transmitted in the contention period and so every terminal knows which ones want to transmit. After the contention period, each terminal that wants to transmit sends its frame in order. An example assuming six terminals has been given in Figure 3.25.

Example 3.4: Analyze the efficiency of bit-map protocols.

At low load, the protocol efficiency is low. For example, at the extreme case, assume one packet transfer per contention period. The efficiency is

$$\frac{d}{d+N}, \tag{3.44}$$

where d is the number of bits in the packet and N is the number of terminals in the network assuming that in each reservation slot, one bit is used. At high load, the protocol efficiency is high. At the extreme case, all terminals have packets to transmit. The efficiency is

$$\frac{Nd}{Nd+N}. \tag{3.45}$$

which is

$$\frac{d}{d+1}. \tag{3.46}$$

Note that while there are many designs of reservation-based protocols, other protocols introduced in this chapter can be seen as special cases of reservation-based protocols. For example, static TDMA/FDMA protocols will reserve fixed time or frequency slots for a new session when it is admitted by the central controller. So the reservation phase is negligible and the data phase is almost infinitely long. The polling period in a poll-based access protocol can be treated as the reservation phase, and so can the token-passing period in token-ring protocols. On the other hand, contention-based protocols, e.g. ALOHA, CSMA, can be treated as reservation protocols without data phases and the reservation packets in reservation phases will instead carry data.

When the reservation and data phases have fixed lengths, the reservation capacity and data capacity can be analyzed. The reservation capacity is defined as the maximum number of reservations the system can handle per time unit and the data capacity as the maximum number of messages or bits successfully delivered per time unit. The smaller of the two limits the system performance. There is a tradeoff between the reservation capacity and data capacity. We need to increase the number of reservation resources to increase the reservation capacity at the expense of slots assigned for data transmission. In most systems, the reservation packets are rather small and the reservation capacity is not usually the bottleneck, especially when the message size is large.

Assume there will be B messages transmitted during a data phase. Assume further that a contention resolution algorithm is used during the reservation phase. In the reservation phase, each reservation slot has a duration δ_c and let N_{res} be the number of slots used in the reservation phase. Assume the network is heavily loaded in the sense that the whole data phase is always occupied by packet transmission. Since the data transmission phase is collision free, the system throughput is achieved using the fraction of time spent in the data transmission mode and is

$$S = \frac{E[B]}{E[B]T_m + \delta_c E[N_{res}]} = \frac{1}{T_m} \frac{1}{1 + \frac{\delta_c E[N_{res}]}{T_m E[B]}}. \tag{3.47}$$

Compared to (3.16), the performance loss is due to the overhead of the reservation phase. There are three main ways of improving the efficiency of reservation-based protocols:

(i) Design the system such that each reservation slot is as short as possible. This is usually limited by network synchronization capability, the necessary amount of information carried in reservation packets, and channel capacity. Note that if δ_c is sufficiently small, the system throughput approaches (3.16) and the time spent in the reservation phase vanishes.

(ii) Design the contention protocol in the reservation phase such that the number of reservation slots, N_{res}, is as small as possible. In general, N_{res} grows with the number of terminals in the network. If only one terminal is allowed to transmit after each reservation phase, it has been shown in [G. W. Miao et al., 2012] that $E[N_{res}]$ can be bounded above by 2.43 even if there is an infinite number of terminals in the network. This is achieved by using slotted ALOHA in the reservation phase and choosing the contention probability of each terminal to be inversely proportional to the number of terminals in the network.

(iii) Allow as many messages transmitted as possible in each data phase. One example of this design is that all terminals succeeding in getting reservations will empty their queues during the data phase. Then the data phase grows in length with λ. With high load, $T_m E[B]$ will be sufficiently high that the system throughput approaches (3.16) and the fraction of time spent in the reservation phase vanishes.

If the network is at low loads, the data phase will be very short. The majority of data phases will, in fact, have zero duration. The typical delay for a single message arriving under these conditions consists of the duration of a reservation phase and the time spent to transmit the actual message. The reservation will take only one reservation slot of duration δ_c. The expected delay becomes

$$D(\lambda \approx 0) \approx T_m + \delta_c. \tag{3.48}$$

Making a more general analysis of a reservation system is complicated and beyond the scope of this text. Readers are referred to [R. Rom and M. Sidi, 1990; G. W. Miao et al., 2012] for more thorough treatments.

3.5 Applications

3.5.1 IEEE 802.11

The IEEE 802.11 Wireless Local Area Network (WLAN), also known as WiFi, can operate in both infrastructure and peer-to-peer modes. In infrastructure mode, WLAN provides connections to wired infrastructure, whereas in peer-to-peer mode, there is no master–slave relationship between terminals and access points. The data delivery of IEEE 802.11 MAC is asynchronous, best-effort and connectionless. The access scheme for 802.11 uses CSMA/CA. A physical carrier sense mechanism is used to determine if the channel is available or not. This is done by measuring the RF energy at the antennas and determining the strength of the received signal. If the signal strength is below a specified threshold the channel is available. The basic access mechanism is illustrated in Figure 3.26. After finding the channel to be busy, the transmission is delayed. The period of time immediately following a busy channel state has the highest probability of collisions. CSMA/CA uses a minimum time gap between frames for each terminal. Once a frame has been sent and the channel goes from the busy state to idle, a terminal must wait until the time gap is up to try a transmission. Once the time has passed, the terminal selects a random amount of time, called the backoff interval, in the contention window (CW) to wait before sensing the channel again to verify if it is still in the idle state. If the channel is busy, another backoff interval is selected with a size that increases in an exponential fashion up to a maximum window size. This process is repeated until the waiting time approaches zero and the terminal is allowed to transmit. This mechanism ensures judicious channel sharing while avoiding collisions.

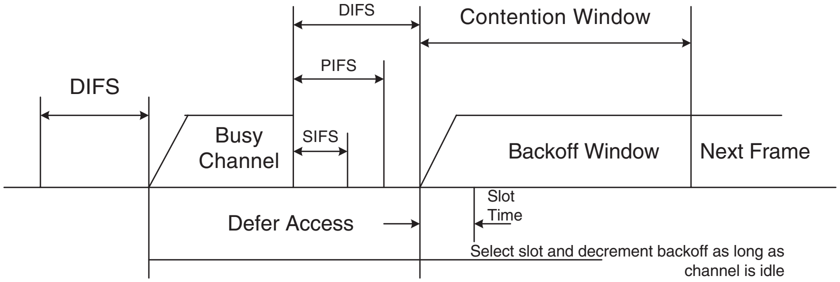


Figure 3.26 IEEE 802.11 basic access protocol.

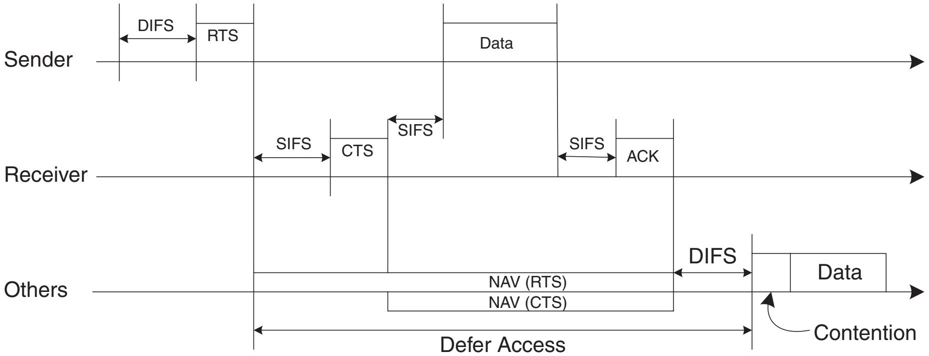


Figure 3.27 IEEE 802.11 RTS/CTS timing.

To further avoid collisions, 802.11 implements additional features. The most important one is a virtual carrier sense mechanism, called the network allocation vector (NAV). The NAV limits the need for physical carrier sensing to save power. The NAV is like a counter. When the counter is zero, the virtual carrier sensing indicates that the medium is idle, and otherwise that it is busy. The channel is determined to be busy when the terminal is transmitting. The NAV maintains a prediction of future traffic on the channel based on duration information that is announced in the network. While the NAV is set, the channel is found to be busy by the protocol. The NAV can be exploited by using RTS, CTS, ACK and data transmission frames sequentially. An example is given in Figure 3.27. A terminal that wants to send data will send an RTS frame, which includes the destination and the length of the message, that can be used to set the NAV in all terminals detecting the RTS frame. This information prevents all terminals from sensing or transmitting for the duration of the message. The receiving terminal issues a CTS frame as a reply. If the CTS frame is not received, a collision has occurred and the RTS process starts over. After a successful data transmission, an ACK is sent back to verify the successfulness. It can be seen that this scheme limits the contention between terminals to only the RTS/CTS frames and the payload data is transmitted in a scheduled and contention-free way.

3.5.2 Cellular networks

The UMTS/3GPP WCDMA system is a conventional cellular system. WLAN is designed to work in free and shared spectrum and therefore it is not highly optimized. It is a simple protocol and easy to implement. The contention time and packet headers in WLAN usually result in significant overhead. On the other hand, WCDMA is designed to operate in licensed spectrum, so it is optimized to maximize the utilization of expensive spectrum. The protocol is therefore much more complex and optimized than WLAN. For packet access in the uplink, two basic modes are designed: packet random access mode using a random access channel (RACH) and dedicated data transmission mode where a dedicated channel (DCH) is allocated for the duration of data transfer.

RACH is a shared channel that is used by mobile terminals to access the base station, especially for initial access. It can also be used for the transport of small and infrequent user data packets. For example, RACH is used in mobile terminals when they need to get the attention of a base station to initially synchronize their transmission with the base station. The RACH is always received by the base station from the entire cell. Its encoding is robust enough to handle varying channel conditions in the cell. The messages transmitted in RACH are not scheduled. When multiple terminals make connections at the same time, collisions may happen. The RACH transmission is based on a slotted ALOHA approach. A mobile terminal can start the random access transmission at the beginning of a number of well-defined time intervals, termed access slots. There are 15 access slots per two frames, as shown in Figure 3.28. Information on which access slots are available for random access transmission is given by higher layers. Some examples of random access transmission are illustrated in Figure 3.28. Several measures can be used to limit the number of collisions. The base station controls the transmission probability p using downlink command signals at scheduled intervals. The base station may allow a subset of terminals, e.g. those of certain service types, to contend for access. Different preamble spreading codes can be used by the terminals so that the RACH is further divided into subchannels. The RACH mode requires no setup time and is ideal for short packets desiring low-delay transmissions. However, each packet has to carry complete information about the terminal and thus creates a large overhead.

The DCH is particularly attractive for longer messages. The reservation of DCH resources for the data transfer is done using a sequence of handshake messages on the RACH and some other control channels. Once the resources are reserved, the terminal can send data on the DCH, using various RRM techniques, e.g. power control, adaptive modulation and coding, interleaving, and so on. The throughput in the DCH is high because explicit terminal information is not needed and the signaling overhead is low. However, as for the performance of reservation-based protocols, there is a tradeoff between the reservation capacity in the control channels, e.g. RACH, and data capacity in DCH. The price paid for the higher DCH throughput is the resource consumption in the other control channels. For example, the time slots in RACH for setting up a data transmission can be of the order of several tens of milliseconds.

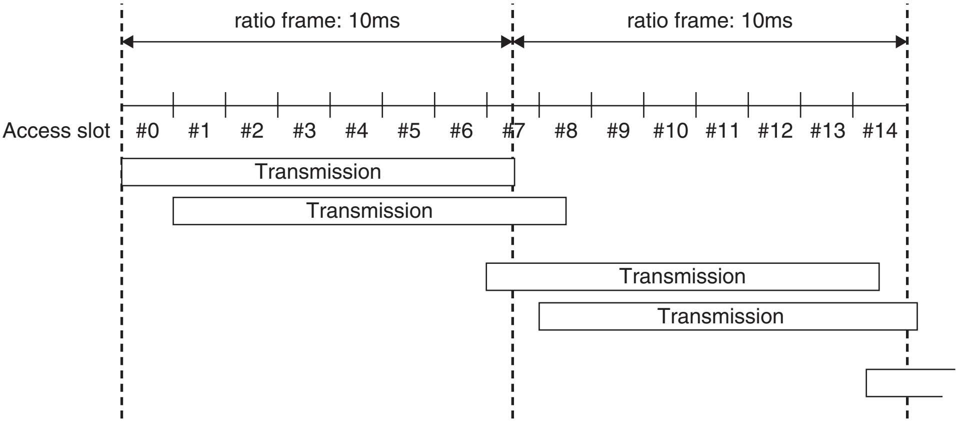


Figure 3.28 RACH access slot numbers and their space [3GPP TS 25.211 V9.2.0, 2010].

Exercises

3.1 Derive the maximum throughput for pure ALOHA.

3.2 A wireless network for taxis uses a radio channel that provides a bit rate of 4800 bps. Each message has 50 characters (8 bits/char). The network employs slotted ALOHA.

a) What is the upper bound of the number of taxis that can be served if each taxi generates a message every minute following the Poisson process.
b) What is the expected packet delay with 100 taxis assuming the average retransmission delay is four seconds.

3.3 In a slotted ALOHA network, assume all collided packets are lost. The transmitted packets are lost with probability P_e due to noise. Assume the compound process of retransmitted and new messages will also be a Poisson process.

a) Determine the maximum throughput.
b) Determine the delay as a function of the throughput for $P_e = 0, 0.05, 0.1, 0.25$. Also plot the functions. Assume $\xi = 20$.

3.4 In a synchronous random access network, four terminals use a conflict resolution algorithm with a binary tree ($q = 2$) to access the system. All terminals are independent of each other and each one has a message to transmit at the beginning of a conflict with probability p.

a) Determine the maximum number of slots needed to resolve a conflict.
b) What is the expected duration of the conflict resolution process in time slots?
c) What is the expected duration of the conflict resolution process assuming the first time slots are consequently skipped ($M_0 = 2$). In order for the system to perform better than the original system, which values should p lie within?

3.5 Consider the uplink of a single-cell slotted ALOHA wireless packet data system. The base station (BS) is placed in the middle of the cell and the users are uniformly distributed over the cell area. The system uses constant received power control and all users will achieve the same data rate, 1024 kbps, if the transmission succeeds. A packet is successfully received at the base station if there is no collision. Each user needs to send 10 kbit messages on average every 10 seconds following the Poisson process, even if the previous messages have not yet been sent.

a) What is the maximum number of users the system can support such that the expected packet delay is finite?
b) If pure ALOHA is used, what is the maximum number of users the system can support?
c) If the BS schedules the transmission of all users in an independent control channel using the round-robin scheduler, what is the maximum number of users the system can support?

References

3GPP TS 25.211 V9.2.0. 2010. 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Physical channels and mapping of transport channels onto physical channels (FDD). Sept.

L. Ahlin, J. Zander and S. Ben Slimane. 2006. *Principles of Wireless Communications*. Lund: Studentlitteratur.

J. J. Capertanakis. 1979. Improvements in Block Retransmission Schemes. *IEEE Trans. Commun.*, 27(Feb.), 524–532.

L. Kleinrock. 1976. *Queueing Systems, Part I: Theory*. New York: John Wiley & Sons.

G. W. Miao, Y. (G.) Li and A. Swami. 2012. Channel aware distributed medium access control. *IEEE/ACM Trans. Networking*, 20(4), 1290–1303.

R. Rom and M. Sidi. 1990. *Multiple Access Protocols*. New York: Springer-Verlag.

4 Scheduling

4.1 Introduction

Wireless communications are evolving from analog, small-capacity, voice services to digital, large-capacity, data services. Nowadays wireless systems should be designed to accommodate many new requirements. For example, wireless networks should be capable of providing high data rates so that terminals can receive broadband services with fast response times. Wireless networks should also have a flexible service architecture to integrate different types of services on a single air interface because terminals have different service requirements. If the network is optimized only for one type of service, other types will experience poor service quality. On top of the flexible service architecture, effective QoS management schemes are also needed. This is because QoS metrics differ among different applications that may even be of the same type. For example, all video telephony has a strict delay requirement but the detailed parameters can be different. When different resolutions of videos are used, the delay requirements of sending each packet would also differ, as would the rate requirements.

The requirements of all terminals can be met easily if there are unlimited wireless resources, e.g. infinite spectrum, infinite transmission power and unlimited antennas, such that each terminal can be allocated whatever resources it desires. In practice this is impossible because of various limitations. The spectrum is allocated by the government and is very limited. Technically it is also difficult to implement devices that support communications over infinite spectrum. The RF transmission power should not exceed government regulations. It is impossible to implement power amplifiers that support infinite power output. In addition there is also the concern of high energy bills. Because of the limits of device dimensions, it is also impossible to use an infinite number of antennas in wireless communications. Therefore, wireless resources need to be shared among all terminals carefully and it is desirable to schedule the usage of wireless resources as efficiently as possible, while maximizing the overall network performance. For example, spectrum bandwidth is a key resource carrying wireless signals and determines the maximum symbol transmission rate. With FDMA, the amount of bandwidth allocated to each terminal limits its channel access rate. In other words, the bandwidth allocation determines the transmission opportunity of each terminal. Similarly, time slots in TDMA and codes in CDMA are all resources that should be scheduled efficiently.

In Chapter 3 we introduced contention-free and contention-based access protocols. With contention-based access protocols, the protocol itself will decide the resource allocation. Consider the pure ALOHA system as an example. Each terminal sends a packet whenever it is generated and if a collision happens, each terminal retries access to the channel later. The protocol itself determines when a transmission should occur and therefore the resource allocation of every terminal in the network. Since the protocol automatically determines the resource allocation, its design is essential in determining the efficiency of resource utilization, as has been demonstrated in Chapter 3. Contention-based access protocols are not efficient as resources are wasted when collisions occur. For example, the typical efficiency of an 802.11a system in the real world is about 50 percent, which means a network with a nominal data rate of 54 Mbps will achieve only a maximum throughput of about 25 Mbps. In spite of potential improvements, it is still difficult to drive the efficiency of contention-based networks beyond 65 percent. Besides, contention-based protocols serve terminals only on a sequential basis as each packet occupies all the network resources. The latency will be huge when there are many terminals in the network. Cellular networks need to serve a large number of terminals with a wide variety of traffic including voice, video and data. Efficiency and low latency are therefore important. Reservation-based access protocols are superior in both of these critical dimensions. Reservation-based access protocols have been briefly introduced in Chapter 3. This type of protocol is the most commonly used access protocol in wireless cellular networks because of its high efficiency as well as flexibility in managing wireless resources. In a reservation-based access protocol with centralized scheduling, the central controller, e.g. the base station in a cellular network, schedules the transmission of all terminals in the network in the reservation phase and all terminals will follow the scheduling results to transmit accordingly in the data phase. An example of using a reservation-based access protocol with centralized scheduling in cellular systems is given in Example 4.1.

Example 4.1: A reservation-based protocol in LTE.

A physical-layer frame structure of LTE is illustrated in Figure 4.1. Each frame consists of 10 sub-frames, each of which has 12 or 14 orthogonal frequency division multiplexing (OFDM) symbols, depending on the length of the cyclic prefix. In each sub-frame, there is a downlink control channel (Physical Downlink Control Channel, PDCCH) and a downlink data channel (Physical Downlink Shared Channel, PDSCH). The PDCCH transmission occurs before the PDSCH transmission. In the example in

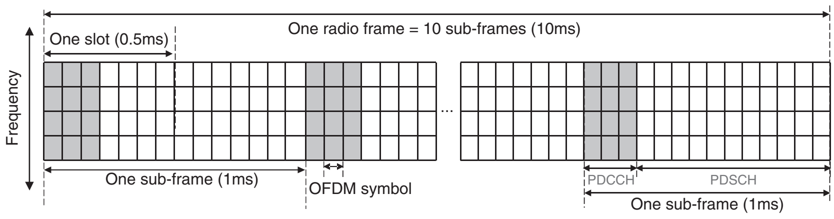


Figure 4.1 A frame structure in LTE.

Figure 4.1, PDCCH uses the first three OFDM symbols and PDSCH the remaining symbols of each sub-frame. The PDCCH conveys control information for each terminal. Therefore robustness rather than maximum data rate is the main consideration and QPSK (quadrature phase shift keying) is the only available modulation format. On the other hand, PDSCH is used basically for data and multimedia transport and therefore designed for very high data rates. Modulation options in PDSCH therefore include QPSK, 16QAM and 64QAM (quadrature amplitude modulation). In addition, PDSCH multiplexes the data of all terminals in the network and the data of each terminal is transmitted on a unique set of resources, i.e. a set of time slots different from all others. The base station schedules the downlink transmission of all terminals and uses PDCCH to reserve the PDSCH resources, e.g. time slots, frequency resource blocks and codes, for each terminal. Therefore a reservation-based protocol is used, in which PDCCH is the reservation phase and PDSCH the data phase.

To get a bigger picture, the role of scheduling in wireless cellular networks is illustrated in Figure 4.2. In each cellular network, each terminal may estimate its channel and feed the result back to the base station. In addition, each terminal may also send to the base station its QoS requirements, such as data rate, delay, buffer status, BER, etc. After collecting the states of channels and QoS requirements of all terminals in the network and checking the queueing state within the base station, the scheduler in the base station will decide the wireless resource allocation for each terminal. The decision result may include time slot, frequency band, antennas, power level, modulation, coding, data rate and so on that are to be used by each terminal for the following data transmission. Then the allocation results will be sent to all terminals. All this happens in the reservation phase. After that the base station will communicate with all the terminals using the predefined wireless resources, which happens in the data phase. Therefore scheduling in wireless networks refers to processes that determine the amount of wireless resources to allocate to each terminal following certain procedures. The goal is to deliver the required service quality for each terminal as far as possible.

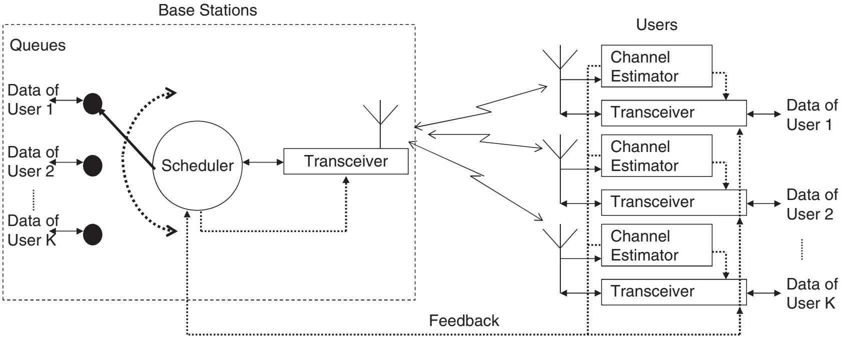


Figure 4.2 Scheduling in cellular networks.

4.2 Issues in wireless scheduling

There are certain factors that affect wireless scheduling: the set of available wireless resources to share among the terminals, the channel information available at the scheduler, the service requirements of each terminal, and the objectives to optimize in relation to the performance metric of the service provider. In this section, we discuss mainly the impact of QoS and wireless channels.

4.2.1 Quality of service

Broadband mobile cellular networks need to support mixed classes of traffic with different QoS requirements. QoS differentiations and guarantees are therefore needed.

Take UMTS as an example, which was discussed in detail in Chapter 2. Here we recap the QoS classes of UMTS and see their impact on wireless scheduling. UMTS defines four QoS classes, conversational, streaming, interactive and background, as shown in Table 2.1. The conversational class has a strict requirement for the transfer time, which should be small enough that the time relation between information entities, i.e. transmitter and receiver, should be preserved. The delay requirement is given by how humans perceive audio or video conversations. The streaming class is the delivery of real-time audio or video services and is a one-way service. This class of services does not require very low transfer delay because it only affects the response in the initial access. Besides, data buffering is usually used on the receiver side to smooth the delay jitter and therefore delay variation can also be larger than for the conversational class. The interactive class characterizes applications like web browsing and remote server access. Here a terminal sends a service request and the remote server sends back the required contents. Terminals receiving this class of service expect to receive the content reliably within a certain time, but not necessarily immediately (not real time). So the delay requirement is much looser than for the conversational and streaming classes. But the content must be preserved, meaning data should be received with very low probability of error. The background class is the least time sensitive and is usually called best-effort service. This class of applications has no time requirements as long as the content can be delivered reliably.

Therefore the traffic in wireless networks may be diverse in rate, delay, BER, service quality and so on. An example is shown in Figure 4.3, where five terminals are accessing the network with different service requirements, including voice, video call, file download, YouTube, and web browsing. Besides requirement differences, terminals may have other limiting factors that differ from each other, e.g. peak power limits. To accommodate these different requirements, wireless networks should be flexible in allocating network resources, which can be achieved using scheduling algorithms. A simple scheduling algorithm is to select terminals according to a fixed cyclic order, yielding a fair transmission opportunity for every terminal. However, if a terminal has urgent packets that need to be delivered within a time limit, i.e. with a delay requirement, this algorithm may fail because the terminal has to wait for its turn. So, priority may need to be considered in the scheduling algorithms. Usually, improvement

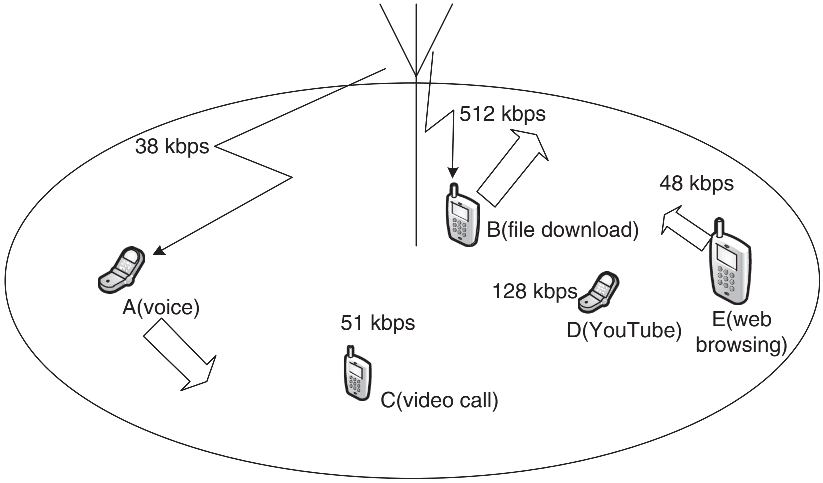


Figure 4.3 A single-cell cellular network (128 kHz system bandwidth).

of the QoS of one terminal will degrade the performance of other terminals. For example, in Figure 4.3, assume only terminals A and B are accessing the network. If the whole system bandwidth is allocated to terminal A, it will achieve 38 kbps throughput; if to B, it will have 512 kbps. Terminal A is receiving voice service and desires only 25 kbps stable data rate while B desires as high a data rate as possible. To obtain the highest network throughput, terminal B should always be scheduled, indicating no service for terminal A at all. Obviously this is not a desirable solution as terminal A will be very unhappy. When there are multiple terminals in the network, all with different delay and rate requirements, like the five terminals in Figure 4.3, the scheduling problem will be even more complicated because the resources are shared by all terminals and it may be difficult to satisfy all terminals at the same time. Therefore, while maximizing system performance, it is also critical for the scheduling to enhance the QoS performance of individual terminals in a balanced way.

4.2.2 Channel variation

Because of shadowing, fading, noise, interference and terminal mobility, the quality of wireless communications is unstable, error prone and cannot be predicted. The capacity of almost every wireless link varies significantly in different time periods and locations. Even when the scheduler knows the QoS requirement of a terminal, it is still difficult to estimate the amount of resources needed to meet the requirement. An adaptive procedure is therefore needed to provide QoS assurance considering terminal requirements and channel variations. Good scheduling approaches should try to exploit channel conditions opportunistically to achieve better network performance. Here the term opportunistic means the ability to schedule terminals based on favorable channel conditions. However, the potential of opportunistic transmissions also introduces a tradeoff between efficiency in using wireless resources and the level of terminal satisfaction. For example, as shown in Figure 4.3, a base station is communicating with several mobile terminals at the same time. The terminals are located at different places and some are very close to the base station, called cell-center terminals, and others far away, called cell-edge terminals. Cell-center terminals can enjoy very reliable communications with the base station, like terminal B, while cell-edge terminals can barely communicate with the base station, like terminal A. Furthermore, terminals may move, as indicated by the arrows in Figure 4.3. This mobility will make the link qualities more dynamic. Therefore the scheduler should have certain dynamic mechanisms to deal with channel variations in different time periods and locations such that the QoS requirements can be met.

4.3 Wireless scheduling and capacity region

Scheduling makes the system flexible enough to adapt to the channel variations and user QoS requirements, thereby permitting a flexible service architecture for integrating various types of services in a single air interface. Scheduling in wireless networks, especially in a cellular mobile network, can lead to significant performance improvement without using more spectrum. This chapter presents a framework for scheduling terminals in an opportunistic way to achieve high network performance by exploiting time-varying channel conditions while assuring the level of QoS among different terminals.

Assume a base station is scheduling the resources of one single channel for M terminals. In each time slot, only one terminal is scheduled. An example is the uplink or downlink communications in cellular networks, as illustrated in Figures 4.4 and 4.5. The arriving traffic of each terminal is stored in an independent queue. At time t, the queue of terminal i changes following the state equation

$$q_i[t+1] = q_i[t] + \delta_i[t] - \eta_i[t], \tag{4.1}$$

where $q_i[t]$ is the number of bits in the queue of terminal i, $\delta_i[t]$ the number of bits arriving at the queue, and $\eta_i[t]$ the number of bits scheduled to transmit in this time slot. Scheduling algorithms are used to determine $\eta_i[t]$ for all terminals in each time slot. To simplify the analysis, the length of each time slot, T, is assumed to be shorter than the channel coherence time and the channel is assumed to be stationary for the duration of each time slot, but may vary from slot to slot. The different queues typically have different QoS constraints, which should be considered by the scheduling algorithms when deciding $\eta_i[t]$. The scheduling should also take into account the achievable data rates of the users, which are determined by the channel qualities and will be illustrated below. The goal of scheduling is to allocate the resource blocks, e.g. time slots, to optimize a certain performance metric, e.g. a function of maximum/minimum/average throughput, delay or outage probability.

The maximum number of bits that can be reliably delivered for all the users in each time slot is bounded above by the so-called multi-user capacity of the channel. When there is only one user in the network, the capacity of a channel, C, is the rate

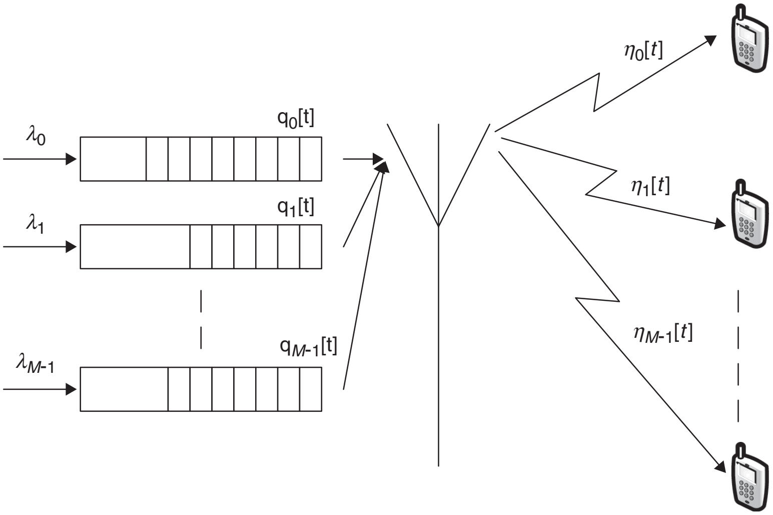


Figure 4.4 Downlink scheduling.

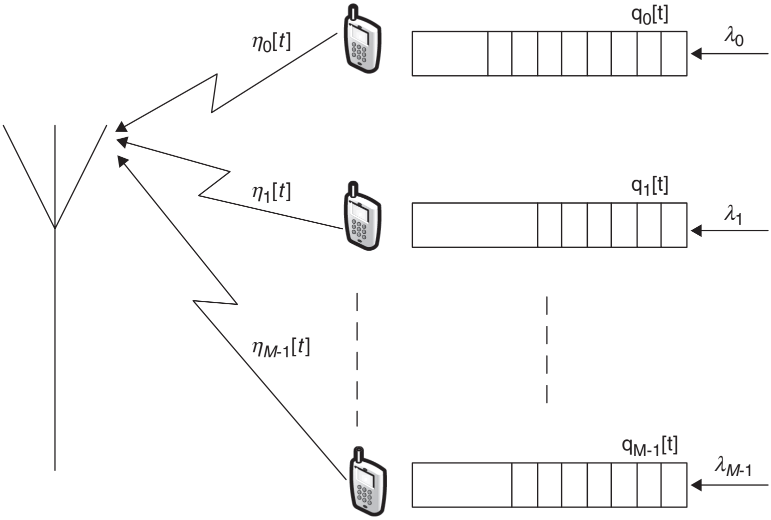


Figure 4.5 Uplink scheduling.

limit for which reliable communications can be attained, i.e. any data rate below C can be achieved with arbitrarily small probability of error. So the user's data rate is bounded above by only one value, C. In the multi-user case, this concept needs to be extended to a capacity region, denoted by $\mathcal{C}$, which is the set of all possible reliable data rates $(r_1, r_2, \ldots, r_M)$ for all users that can be achieved simultaneously. The points on the outer boundary of the capacity region represent the best data rates that can be achieved using the optimal multi-user code designs and certain scheduling algorithms. The interior points represent the data rates obtained using suboptimal coding schemes. Any point outside the capacity region means at least one user having an error probability bounded away from zero regardless of the coding scheme used. Since all users share the same radio resource, there is a tradeoff between the data rates of the users. If one user communicates at a higher data rate, the other users may need to lower their data rates. The scheduling algorithms can be employed to determine the data rate of each user while assuring the QoS of all users. In the following we focus on additive white Gaussian noise (AWGN) channels without fading to illustrate the multi-user capacity region and its impact on wireless scheduling in detail. More discussions about multi-user capacity can be found in [D. Tse and P. Viswanath, 2005].

4.3.1 Uplink multi-user capacity

Consider the uplink transmission with two users whose channels are static with zero-mean additive white Gaussian noise. The capacity region is the set of all rates (r_1, r_2) that satisfy the following three constraints:

$$\begin{aligned} r_1 &\le W \log_2\left(1+\frac{P_1 g_{10}}{N_0 W}\right) \\ r_2 &\le W \log_2\left(1+\frac{P_2 g_{20}}{N_0 W}\right) \\ r_1 + r_2 &\le W \log_2\left(1+\frac{P_1 g_{10}+P_2 g_{20}}{N_0 W}\right), \end{aligned} \tag{4.2}$$

where W is the frequency bandwidth, P_i the transmission power of user i, g_{i0} the power gain of the channel of user i, and N_0 the noise spectral density. The first two constraints are from the capacity limit of each individual link and the third corresponds to the sum capacity of the two-user uplink channel. Figure 4.6 illustrates this capacity region, which has a pentagon shape. At point A, the first user should transmit at data rate $W \log_2(1+\frac{P_1 g_{10}}{N_0 W})$ while the second user is silent. Similarly at point D, user 2 will transmit at $W \log_2(1+\frac{P_2 g_{20}}{N_0 W})$ while user 1 is shut down. Point C can be achieved by letting user 2 transmit at the full data rate $W \log_2(1+\frac{P_2 g_{20}}{N_0 W})$ with user 1 transmitting at a data rate that treats the signals from user 2 as interference. To receive data from the two users, the so-called successive interference cancelation (SIC) receiver needs to be implemented at the base station. The SIC receiver first detects the signals from user 1 while treating the signals from user 2 as interference. Then the receiver reconstructs the received signal from user 1 and subtracts it from the aggregated received signal. Finally the receiver can decode the data of user 2. The other corner point, B, can be achieved by reversing the cancelation order. The other points on the line BC can be achieved by time-sharing between the transmission strategies in points B and C and a scheduling policy may be used to determine how the time is shared.

In general, when there are M uplink users, and the sum data rate is bounded above by

$$\sum_{i=1}^{M} r_i \le W \log_2\left(1+\frac{\sum_i P_i g_{i0}}{N_0 W}\right), \tag{4.3}$$

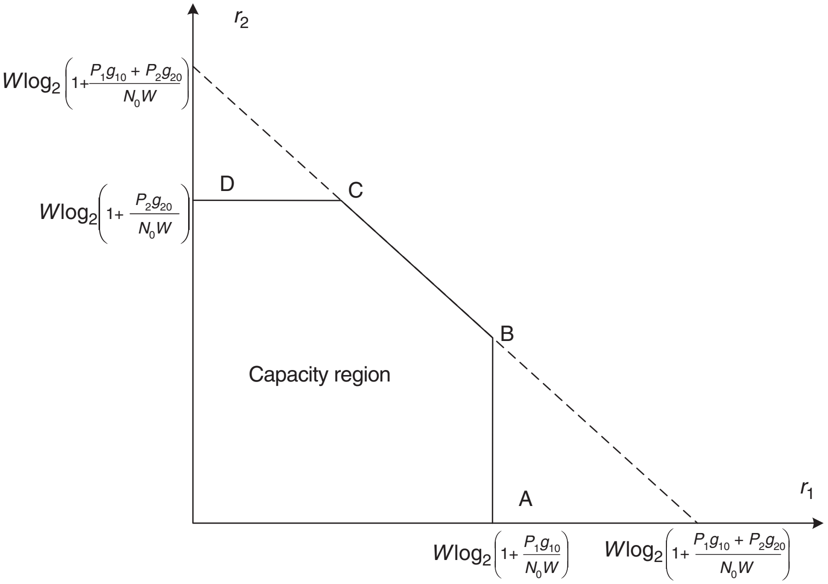

r_2
$W\log_2\left(1+\frac{P_1g_{10}+P_2g_{20}}{N_0W}\right)$
$W\log_2\left(1+\frac{P_2g_{20}}{N_0W}\right)$
D
C
B
Capacity region
A
r_1
$W\log_2\left(1+\frac{P_1g_{10}}{N_0W}\right)$
$W\log_2\left(1+\frac{P_1g_{10}+P_2g_{20}}{N_0W}\right)$

Figure 4.6 Capacity region of the uplink with two users.

the corresponding capacity region can be obtained by induction. For example, with three users, the capacity region is

$$\begin{aligned}
r_1 &\le W\log_2\left(1+\frac{P_1g_{10}}{N_0W}\right)\\
r_2 &\le W\log_2\left(1+\frac{P_2g_{20}}{N_0W}\right)\\
r_3 &\le W\log_2\left(1+\frac{P_3g_{30}}{N_0W}\right)\\
r_1+r_2 &\le W\log_2\left(1+\frac{P_1g_{10}+P_2g_{20}}{N_0W}\right) \qquad (4.4)\\
r_2+r_3 &\le W\log_2\left(1+\frac{P_2g_{20}+P_3g_{30}}{N_0W}\right)\\
r_1+r_3 &\le W\log_2\left(1+\frac{P_1g_{10}+P_3g_{30}}{N_0W}\right)\\
r_1+r_2+r_3 &\le W\log_2\left(1+\frac{P_1g_{10}+P_2g_{20}+P_3g_{30}}{N_0W}\right),
\end{aligned}$$

which is illustrated in Figure 4.7.

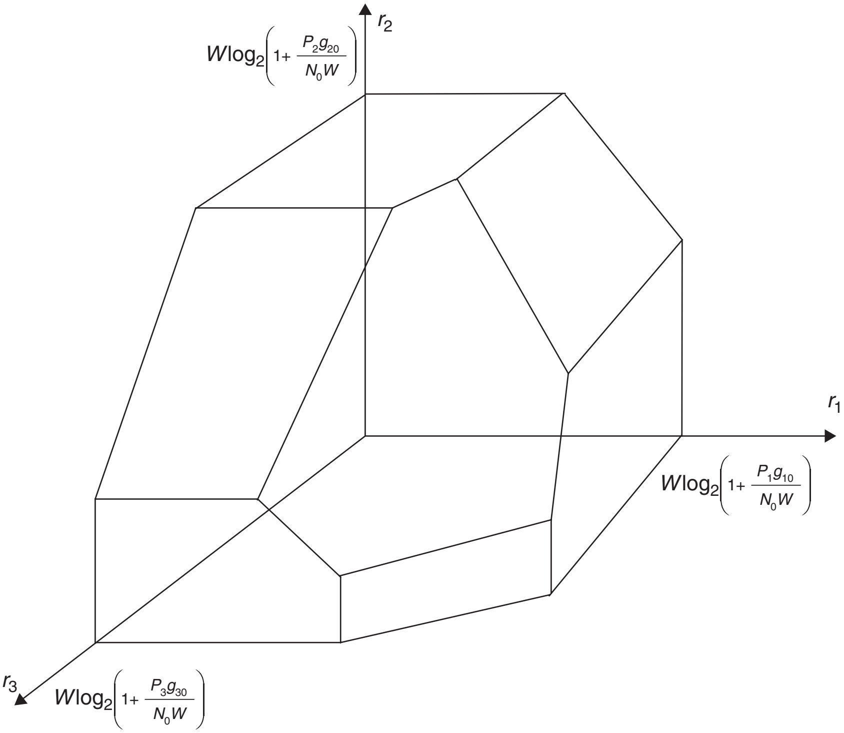

r_2
$W\log_2\left(1+\frac{P_2g_{20}}{N_0W}\right)$
r_1
$W\log_2\left(1+\frac{P_1g_{10}}{N_0W}\right)$
r_3
$W\log_2\left(1+\frac{P_3g_{30}}{N_0W}\right)$

Figure 4.7 Capacity region of the uplink with three users.

4.3.2 Downlink multi-user capacity

In the downlink, the base station sends independent data streams to multiple users. Assume $g_{01} \le g_{02} \le \cdots \le g_{0M}$. The capacity region is given by

$$r_m \le W\log_2\left(1+\frac{P_m g_{0m}}{\sum_{i=m+1}^{M} P_i g_{0m} + N_0 W}\right), \qquad \forall m \tag{4.5}$$

for all possible power allocations $\sum_m P_m = P_0$, where P_0 is the total transmission power of the BS. To achieve optimal data rates, the BS needs to employ so-called superposition coding and the users need to apply SIC receivers. The main idea is to superpose the data of better users on those of the poorer users; each better user always first decodes the data of the poorer users and then its own data. More information about the superposition coding and SIC can be found in [D. Tse and P. Viswanath, 2005].

An example of the capacity region of the downlink transmission with two users is shown in Figure 4.8. Note that the two corner points correspond to the full power allocation to either user 1 or user 2.

It is shown in [N. Jindal et al., 2004] that the capacity region of the downlink broadcast channel is equal to the union of the capacity regions of the dual uplink multiple access channel with the sum power constraint $\sum_{i=1}^{M} P_i = P_0$, which is illustrated in Figure 4.9 assuming there are two users. In other words, if we assume there

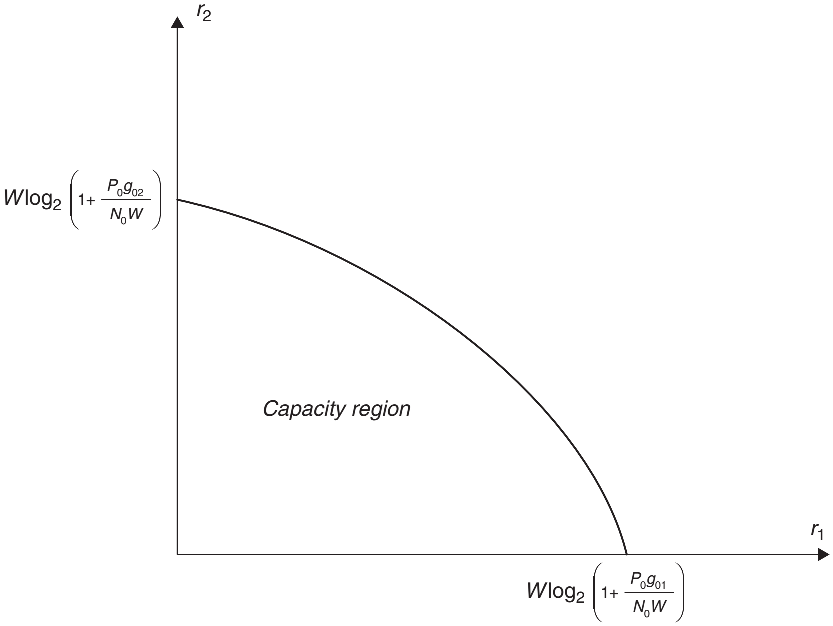


Figure 4.8 Capacity region of the downlink with two users.

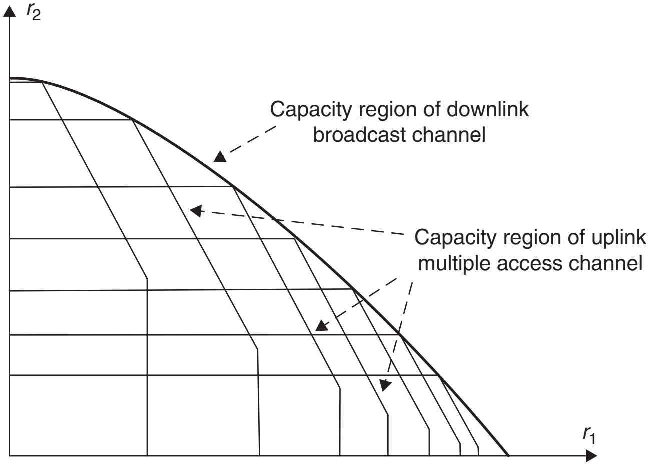

r_2
Capacity region of downlink broadcast channel
Capacity region of uplink multiple access channel
r_1

Figure 4.9 Duality of the capacity regions.

is the same sum power constraint in the uplink, $\sum_{i=1}^{M} P_i = P_0$, instead of the individual power constraints, P_i, the two capacity regions are equal. So the only difference between the uplink and the downlink is the individual power constraints of the users in the uplink transmission.

For both downlink and uplink communications, each point within the capacity region corresponds to a certain coding scheme and resource allocation result that is determined by the scheduling algorithm. As different users might have different QoS requirements, the overall goal of coding and scheduling is to determine a certain point in the capacity region to meet the performance requirements of as many users as possible. A joint design of the coding and scheduling may achieve the globally optimal result, i.e. a point on the boundary of the capacity region, but usually has very high complexities. In addition, the channel usually varies in different frequencies and time slots and it is usually difficult to determine the capacity region and the corresponding optimal schemes. To simplify the design, coding and scheduling are usually implemented separately. In this chapter, the focus is on the scheduling aspect, assuming the coding has been given. In addition, orthogonal coding is assumed and in each resource unit, e.g. a time slot, only one user will be scheduled. While achieving suboptimal performance, i.e. a point inside the capacity region, the implementation complexity is negligible, as will be demonstrated in the following sections.

4.4 Round-robin scheduling

Round-robin (RR) scheduling is one of the simplest scheduling algorithms. Terminals in a round-robin algorithm are scheduled in a round-robin, i.e. cyclic order, manner. Mathematically, the index, $i[t+1]$, of the terminal that is scheduled at time $t+1$ is given by

$$i[t+1] = i[t] + 1. \tag{4.6}$$

It can be seen that the scheduler always selects the terminal that has not been served for the longest time. RR scheduling is fair in the sense that it gives all terminals the same amount of time resources. However, this algorithm does not use the terminal channel quality information and may suffer from low throughput. In the following we introduce more advanced schedulers to improve network performance. These schedulers are frequently referred to as opportunistic scheduling or channel-aware scheduling algorithms as they exploit the channel statistics to improve network performance.

4.5 Max throughput scheduling

The goal of max throughput scheduling is to schedule the terminal in each time slot such that the total network throughput is maximized. If terminal i is scheduled, the expected instantaneous throughput in this slot would be

$$\hat{r}_i[t] = \frac{\hat{\eta}_i[t]}{T_s}, \tag{4.7}$$

where T_s is the slot length and $\hat{\eta}_i[t]$ is the estimated number of bits that can be successfully delivered. The total expected network throughput would be

$$\hat{r}[t] = \sum_{i=0}^{M-1} \hat{r}_i[t] I(i), \tag{4.8}$$

where $I(i)$ denotes the scheduling indicator and is 1 if terminal i is scheduled and 0 otherwise. The goal is to schedule the terminal such that the total network throughput is maximized. Therefore, the terminal with the highest expected throughput should be scheduled to maximize $\hat{r}[t]$.

One way of estimating $\hat{r}_i[t]$ is

$$\hat{r}_i[t] = W \log_2 \left(1 + \frac{\Gamma_i[t]}{\theta} \right), \tag{4.9}$$

where W is the frequency bandwidth and θ the SINR gap that defines the gap between the channel capacity and a practical coding and modulation scheme. Note that in the downlink, the power allocation is constrained by the total transmission power of the base station, while in the uplink it is constrained by inter-cell interference and the power available at the terminals. While a global optimal design requires joint optimization of power allocation and scheduling, we focus only on the scheduling issue in this chapter. So $\Gamma_i[t]$ is the SINR at time t given the allocated power. The power allocation issue will be discussed separately in Chapter 6.

Obviously, the terminal with the highest estimated throughput $\hat{r}_i[t]$ is also the terminal with the highest SINR, $\Gamma_i[t]$. Therefore the max throughput scheduler is also frequently called the max SINR scheduler or max C/I (carrier to interference) scheduler. Since the variations of channel condition are used for scheduling, the max SINR scheduler is a kind of channel-aware scheduler.

The max throughput scheduler is the most aggressive advanced packet scheduler, which always schedules the terminal with the best instantaneous channel quality. While maximizing network throughput, the main drawbacks are the unfairness and coverage limitations. With this scheduler, the terminals with the most favorable positions or channels will have the highest throughput, but terminals in unfavorable positions may never be served. If there are many terminals in the network, most of them will suffer from scheduling starvation because they have to wait until the terminal with the best channel has no more data to transfer and no other terminals with better channels are admitted.

Example 4.2: Consider a downlink wireless system with two terminals where max throughput scheduling is applied. Each terminal experiences Rayleigh fading.
a) Derive the probability for a terminal being scheduled when its SINR is Γ_i.
b) Assume the rate is $\hat{r}_i = \beta \Gamma_i$. Derive the expected throughput for all terminals.

Solution: a) The ith terminal is scheduled when its SINR is higher than the other one j and

$$\begin{aligned} \Pr(\textit{terminal i is scheduled}) &= \Pr(\Gamma_i > \Gamma_j) \\ &= 1 - \exp\left(-\frac{\Gamma_i}{\overline{\Gamma}_j} \right), \end{aligned} \tag{4.10}$$

where $\overline{\Gamma}_j$ is the average SINR of terminal j.
b) The average throughput of terminal i is

$$\begin{aligned}
E\{S_i\} &= E\left\{\hat{r}_i \Pr(\textit{terminal i is scheduled with } \Gamma_i)\right\} \\
&= E\left\{\beta\Gamma_i\left(1-\exp\left(-\frac{\Gamma_i}{\overline{\Gamma}_j}\right)\right)\right\} \\
&= \int_0^\infty \left\{\beta\Gamma_i\left(1-\exp\left(-\frac{\Gamma_i}{\overline{\Gamma}_j}\right)\right)\right\}\frac{1}{\overline{\Gamma}_i}\exp\left(-\frac{\Gamma_i}{\overline{\Gamma}_i}\right)d\Gamma_i \\
&= \beta\overline{\Gamma}_i - \beta\overline{\Gamma}_i\left(\frac{\overline{\Gamma}_j}{\overline{\Gamma}_i+\overline{\Gamma}_j}\right)^2.
\end{aligned} \tag{4.11}$$

Similarly, the average throughput of terminal j is

$$E\left\{S_j\right\} = \beta\overline{\Gamma}_j - \beta\overline{\Gamma}_j\left(\frac{\overline{\Gamma}_i}{\overline{\Gamma}_i+\overline{\Gamma}_j}\right)^2. \tag{4.12}$$

4.6 Proportional fair scheduling

Proportional fair (PF) scheduling is a compromise scheduling policy, trying to balance the competing interests of maximizing total network throughput and providing all terminals with at least a minimal level of service. The objective of the proportional fair scheduler is to maximize

$$\sum_{i=0}^{M-1} \ln S_i, \tag{4.13}$$

where S_i is the long-run throughput of terminal i. S_i may change in every time slot; define the throughput in slot $t-1$ to be $S_i[t-1]$. $S_i[t]$ can be predicted using an exponential low pass filter,

$$\hat{S}_i[t] = (1-\frac{1}{\tau})S_i[t-1] + \frac{1}{\tau}\hat{r}_i[t]I(i), \tag{4.14}$$

where $\tau >> 1$. It can be shown that in order to maximize $\sum_{i=0}^{M-1}\ln S_i$, the terminal with the highest

$$\frac{\hat{r}_i[t]}{S_i[t-1]} \tag{4.15}$$

should be scheduled.

The reason that the PF scheduler is proportional fair is because it meets the proportional fairness criterion. A feasible vector of throughputs $\mathbf{x} = [x_1, x_2, \ldots, x_n]$ is proportionally fair if for any other feasible vector $\hat{\mathbf{x}}$, the sum of proportional changes is non-positive, i.e.

$$\sum_{i=1}^{n} \frac{\hat{x}_i - x_i}{x_i} \leq 0. \tag{4.16}$$

When a network has already been proportional fair and we change the scheduler such that the throughput of one terminal is increased by a%, there will be a more than a% cumulative decrease of the throughputs of all other terminals.

Now let's seek an intuitive understanding of why the objective in (4.13) is indeed proportional fair. Assume $\delta_i \ll x_i$. Then a feasible vector $\mathbf{x}$ satisfies

$$\begin{aligned}\sum_{i=1}^{n} \ln(x_i + \delta_i) &= \sum_{i=1}^{n} \ln x_i + \sum_{i=1}^{n} \ln(1 + \frac{\delta_i}{x_i}) \\ &\approx \sum_{i=1}^{n} \ln x_i + \sum_{i=1}^{n} \frac{\delta_i}{x_i} \\ &\leq \sum_{i=1}^{n} \ln x_i.\end{aligned} \tag{4.17}$$

Therefore, a proportional fair throughput vector is a vector that maximizes $\sum_{i=1}^{n} \ln x_i$.

Example 4.3: Two terminals are scheduled by the PF scheduler in a TDMA wireless network. The rates of the two terminals are 32 and 128 kbps. The transmission continues for 12 time slots. Assume $S_1[1] = 32$ kbps and $S_2[1] = 128$ kbps. The average throughput is updated using

$$S_i[t] = \frac{t-1}{t} S_i[t-1] + \frac{1}{t} \hat{r}_i[t].$$

Calculate the average throughput of each terminal and for how many time slots each terminal will be scheduled.

Solution: In the first slot,

$$i[1] = \arg\max_i \left(\frac{32}{32}, \frac{128}{128} \right).$$

Since both terminals have the same scheduling metric, 1, either terminal can be selected. Suppose in this case terminal 1 is selected. Then

$$S_1[2] = 1/2 \times S_1[1] + 1/2 \times \hat{r}_1[2] = 1/2 \times 32 + 1/2 \times 32 = 32 \text{ kbps};$$

and

$$S_2[2] = 1/2 \times 128 + 0 = 64 \text{ kbps}.$$

In the second slot,

$$i[2] = \arg\max_i \left(\frac{32}{32}, \frac{128}{64} \right) = 2.$$

Terminal 2 is selected and

$$S_1[3] = 2/3 \times 32 + 0 = 21.33\texttt{kbps};$$

and

$$S_2[3] = 2/3 \times 64 + 1/3 \times 128 = 85.33 \text{ kbps}.$$

In the third slot,

$$i[3] = \arg\max_i \left(\frac{32}{21.33}, \frac{128}{85.33} \right) = 1.$$

Terminal 1 is selected and

$$S_1[4] = 3/4 \times 21.33 + 1/4 \times 32 = 24 \text{ kbps};$$

and

$$S_2[4] = 3/4 \times 85.33 + 0 = 64 \text{ kbps}.$$

This process is repeated until the 12th slot and

$$S_1[13] = 17.2 \text{ kbps};$$

and

$$S_2[13] = 68.9 \text{ kbps}.$$

The terminals are selected in turn and each one is assigned six slots in total.

4.7 Max–min scheduling

The objective of a max–min scheduler is to maximize the minimum performance of all terminals in the network, i.e.

$$\max \min_i S_i. \tag{4.18}$$

The scheduler result is max–min fair if and only if a further increase of throughput of one terminal will result in the decrease of a terminal that already has a smaller throughput.

This can be understood with the following example. There are M empty cylindrical buckets (terminals), all with the same radius but different heights. We need to allocate a certain amount of water (resource) to the buckets. With max–min allocation, any small amount of water should be distributed equally among all the buckets that are not full yet. Repeat this process until either all water is allocated or all buckets are full. Figure 4.10 illustrates an allocation result and we can see that buckets will have the same amount of water if they are not full yet. Besides, if we want to increase the water in a bucket that is not full yet, the water in another bucket that has less or an equal amount will decrease.

Now let’s see who should be scheduled. Combining (4.14) and (4.18), we have the following equivalent objective,

$$\max \min_i (1 - \frac{1}{\tau}) S_i[t-1] + \frac{1}{\tau} \hat{r}_i[t] I(i). \tag{4.19}$$

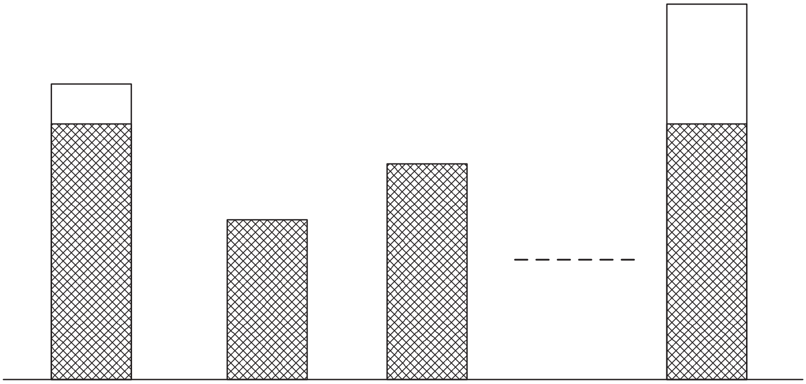


Figure 4.10 Water allocation achieving max–min fairness.

Therefore, we should schedule the terminal with the minimum

$$(1-\frac{1}{\tau})S_i[t-1], \tag{4.20}$$

that is, the one with the lowest throughput at time $t-1$.

Example 4.4: Consider a wireless TDMA system with five terminals accessing the base station. The data rates of the terminals are 1 Mbps 2 Mbps, 4 Mbps, 10 Mbps and 16 Mbps respectively. Compute the throughput of each terminal and the whole network when max–min fair scheduling is used.

Solution: The scheduler assigns the time slots to the terminals so that they achieve the same throughput. Assume terminal i is assigned x_i percent of the slots. Then

$$x_i r_i = x_j r_j.$$

Further, we have $\sum_i x_i = 1$. Solving these equations we have

$$x_1 = \frac{80}{153}, x_2 = \frac{40}{153}, x_3 = \frac{20}{153}, x_4 = \frac{8}{153} \text{ and } x_5 = \frac{5}{153}.$$

Each terminal has throughput $S_i = 1 \times \frac{80}{153} = 0.523$ Mbps. The network throughput is $S = 2.61$ Mbps.

4.8 Max utility scheduling

The previous schedulers are not related explicitly to the QoS that terminals desire to get, nor assume all terminals are demanding best-effort data services. To fill this gap, utility-based scheduling algorithms can be used. Utility quantifies the satisfaction of each terminal given the allocated resources. With utility-based scheduling, the objective is to maximize the utility sum of all terminals, i.e. the total network satisfaction.

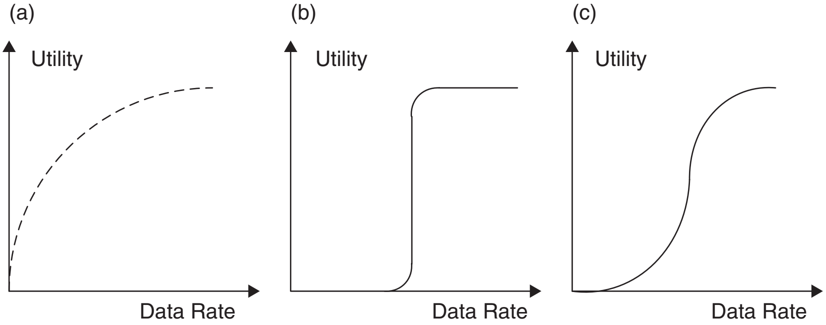


Figure 4.11 Utility of various traffic types with respect to data rate.

To model the relation between services provided by the network and how terminals perceive the services, utility functions can be used. Utility functions should be determined based on traffic characteristics. Figure 4.11 illustrates three popular utility functions. The first one in (a) can be used for best-effort data traffic, e.g. file transfer and email. These services are elastic and terminals can adapt well to long delays and low throughput. On the other hand, more utility can be achieved if a higher data rate can be achieved, but the increase in utility slows with respect to increased data rate. This type of utility function can be used for the interactive and background classes of services in Table 2.1. The type of utility function in sub-figure (b) can be used to characterize real-time services with strict delay requirements, e.g. the conversational class. These services will not be acceptable if packet delays are beyond a certain limit or the rate is smaller than a certain amount. On the other hand, a higher data rate will not help either because these applications have an almost constant rate requirement and higher allocated bandwidth will not be used. Therefore the utility function behaves like a step function and the utility increases from zero to the saturation point quickly beyond the desired rate. The utility function in sub-figure (c) can be used to describe real-time services with looser delay requirements, e.g. the streaming class. For example, streaming video usually uses some scalable coding techniques and can therefore adapt to packet delay and data rate. For example, these applications may be able to adjust the service quality, e.g. video resolution, based on channel conditions and therefore have elastic rate requirements. However, if the channel rate is smaller than the minimum requirement, the performance would be extremely poor, as terminals may not be able to receive any service at all. While increasing the channel data rate will improve the service quality and therefore the utility, there is an intrinsic upper limit of the rate requirement. Therefore when the data rate provided goes beyond the upper limit, the utility saturates and the additional gain diminishes very fast.

Utility-based scheduling aims to maximize the utility sum of all terminals in the network, that is,

$$\max \sum_{i=0}^{M-1} U_i, \tag{4.21}$$

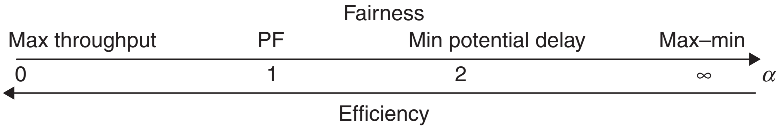


Figure 4.12 Alpha fair scheduling.

where U_i is the utility of terminal i given the allocated resources. Here we focus on the throughput performance and want to find a scheduling result that maximizes

$$\max \sum_{i=0}^{M-1} U_i(S_i). \tag{4.22}$$

Different utility functions can be designed. For example, if

$$U(S) = S, \tag{4.23}$$

the max utility scheduler is the max throughput scheduler. If

$$U(S) = \ln(S), \tag{4.24}$$

the max utility scheduler is the PF scheduler. A more generic definition of the utility function can be [J. Mo and J. Walrand, 2000]

$$U_\alpha(S) = \begin{cases} \frac{S^{1-\alpha}}{1-\alpha} & \alpha \geq 0 \text{ and } \neq 1 \\ \ln(S) & \alpha = 1. \end{cases} \tag{4.25}$$

The parameter α measures how fair the scheduler is (see Figure 4.12): the total network throughput is maximized when $\alpha = 0$, which is the max throughput scheduler, but the resource allocation is not fair at all. When $\alpha = 1$, it is the proportional fair scheduler. When $\alpha = 2$, the goal is to minimize

$$\sum_{i=0}^{M-1} \frac{1}{S_i}. \tag{4.26}$$

Note that the potential delay in transferring packets equals $\frac{1}{S_i}$ for terminal i. So the objective in (4.26) seeks a resource allocation to minimize the total potential delay of all terminals. In the limiting case where $\alpha = \infty$, it is the most fair allocation, the max–min scheduler.

Curves of $U_\alpha(S)$ when α has different values are drawn in Figure 4.13. It can be easily seen that the utility functions in (4.23), (4.24) and (4.25) are the detailed implementations of the type of utility for best-effort data traffic in sub-figure (a) of Figure 4.11. Indeed, sub-figure (a) of Figure 4.11 has a concave monotonic increasing shape and so is $U_\alpha(S)$ for all α values. Other types of utility functions can also be defined, following the shapes in Figure 4.11, to reflect how terminals perceive the QoS and therefore assure QoS in scheduling. Different applications may have different utility functions or even different parameters. For instance, utility functions can also be defined with respect to delay, instantaneous data rate $r[t]$ and so on, and the detailed scheduling

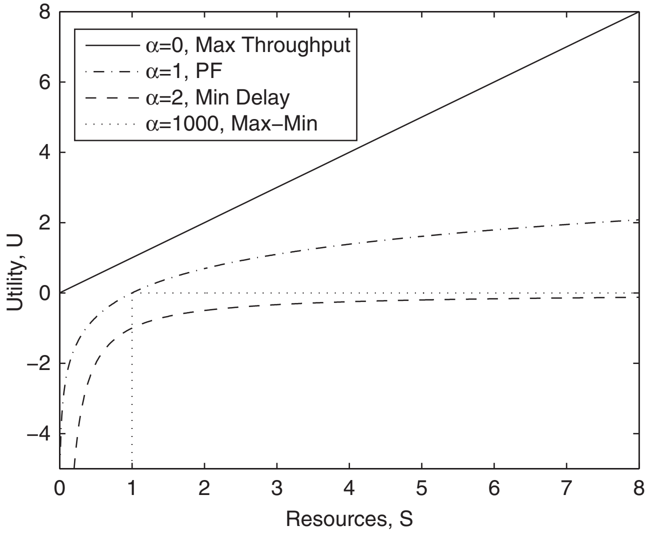


Figure 4.13 Alpha fair utility.

algorithm may vary accordingly. The reader may refer to [B. G. Lee et al., 2009] for more discussion. The focus here is on utility functions with respect to the long-run throughput, the same as other schedulers introduced in this chapter.

Now let's see how to schedule terminals. We want

$$\begin{aligned} &\max \sum_{i=0}^{M-1} U_i(\hat{S}_i[t]) \\ &= \max \sum_{i=0}^{M-1} U_i\left((1-\frac{1}{\tau})S_i[t-1] + \frac{1}{\tau}\hat{r}_i[t]I(i) \right). \end{aligned} \tag{4.27}$$

Note that $(1-\frac{1}{\tau})S_i[t-1] \gg \frac{1}{\tau}\hat{r}_i[t]I(i)$ and we have the following approximation:

$$\begin{aligned} &U_i\left((1-\frac{1}{\tau})S_i[t-1] + \frac{1}{\tau}\hat{r}_i[t]I(i) \right) \\ &\approx U_i\left((1-\frac{1}{\tau})S_i[t-1] \right) \\ &+ U_i'\left((1-\frac{1}{\tau})S_i[t-1] \right) \frac{1}{\tau}\hat{r}_i[t]I(i), \end{aligned} \tag{4.28}$$

where the first portion is fixed at time t regardless of the scheduling result. The objective can be approximated by

$$\max \sum_{i=0}^{M-1} U_i'\left((1-\frac{1}{\tau})S_i[t-1] \right) \frac{1}{\tau}\hat{r}_i[t]I(i). \tag{4.29}$$

Table 4.1 QoS requests

Traffic	Type	Basic rate requirement (kbps)
High rate file download	Best effort	1740
VoIP (voice over IP)	Real time	102
Video streaming	Real time	580

Table 4.2 Performance

Scheduler	Throughput (Mbps)	Outage probability (%)
Max SINR	17.4	5.9
RR	9.2	3.0
PF	13.0	2.1
Utility	10.9	0

Therefore the optimal scheduling policy is to schedule the terminal with the largest

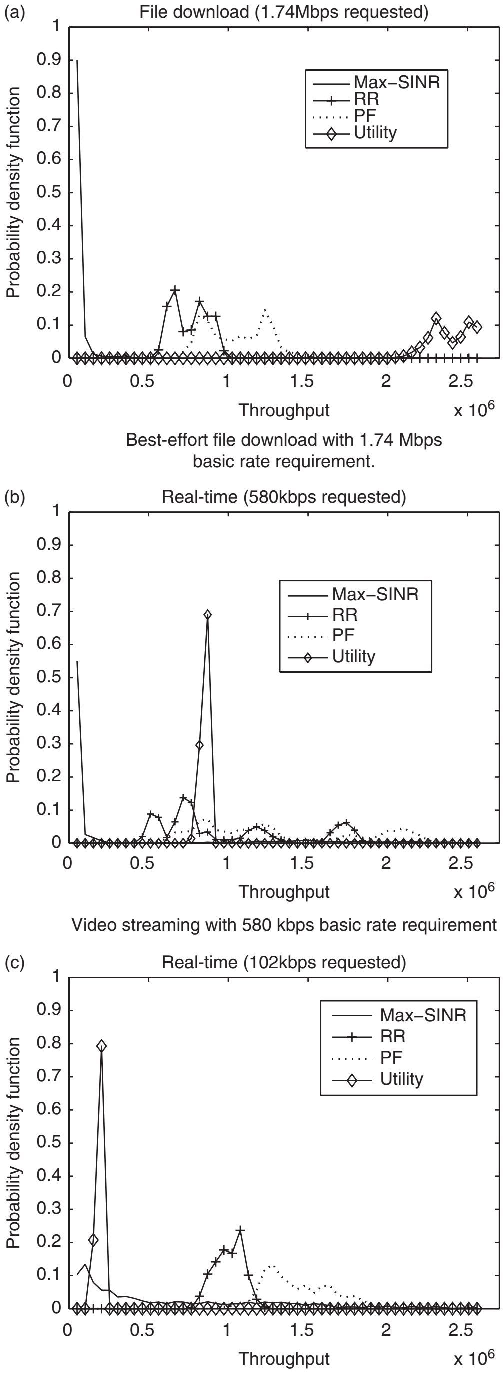
$$U_i'\left((1-\frac{1}{\tau})S_i[t-1]\right)\hat{r}_i[t]. \tag{4.30}$$

Or, observing that $\frac{1}{\tau} \to 0$, the terminal with the largest

$$U_i'\left(S_i[t-1]\right)\hat{r}_i[t]. \tag{4.31}$$

These two perform almost the same.

In the following we show how utility functions can be used to assure QoS for terminals and improve the efficiency of resource utilization. Figure 4.14 illustrates the probability of assigning a specific throughput across each service class in a simulated cellular network [G. W. Miao and N. Himayat, 2008]. Here three service classes are considered, as listed in Table 4.1. Utility functions designed in [G. W. Miao and Z. Niu, 2006] are used. Table 4.2 further compares the corresponding average throughput and outage probability of the network. If the basic data rate requirement of one terminal is not satisfied, we say that the terminal is in outage. It can be seen that the throughput achieved using other schedulers, i.e. max SINR, RR and PF, does not relate to the QoS requirements, as shown in the three figures. On the other hand, the utility-based scheduler is able to concentrate the resource allocation around the rates required for yielding high satisfaction for each service class. Thus they avoid unnecessary allocation to terminals who cannot derive any benefit from the extra resources assigned. The utility-based scheduler achieves 0% outage probability, indicating that given a network outage requirement, the system is able to satisfy the QoS requirements of a larger number of terminals through efficient allocation of resources.



Best-effort file download with 1.74 Mbps basic rate requirement.

Video streaming with 580 kbps basic rate requirement

VoIP with 102 kbps basic rate requirement.

Figure 4.14 Probability density of terminal throughput of different applications: each cell has ten terminals randomly dropped with a random traffic type.

4.9 Scheduling in OFDMA systems

In this section, we introduce scheduling in OFDMA systems. Compared to the scheduling algorithms in the previous sections, there is one more dimension of freedom, subcarrier allocation. Different terminals experience independent wireless channels and their subcarriers may experience substantially different channel gains because of the frequency selectivity in the channels. Theoretically, it is possible to set the data rate for each subcarrier based on its channel quality and the power allocated. Subcarriers in deep fading for one terminal may not be used by this terminal as sending bits on these subcarriers consumes too much power. These subcarriers may be in good conditions on the channels of other terminals and can be used by them instead. This motivates the scheduling of subcarriers in an adaptive way based on the instantaneous channel qualities. OFDMA systems typically use adaptive subcarrier assignment, power allocation, modulation and coding to exploit the diversity in multiple terminals and frequency to improve the network performance. Therefore the scheduling in OFDMA involves subcarrier allocation, link adaptation and QoS assurance of all terminals. It is usually much more complicated than the scheduling introduced in the previous sections. Figure 4.15 illustrates the downlink scheduling of an OFDMA system that is considered in this section. The uplink scheduling can be similarly derived. There are *M* terminals to

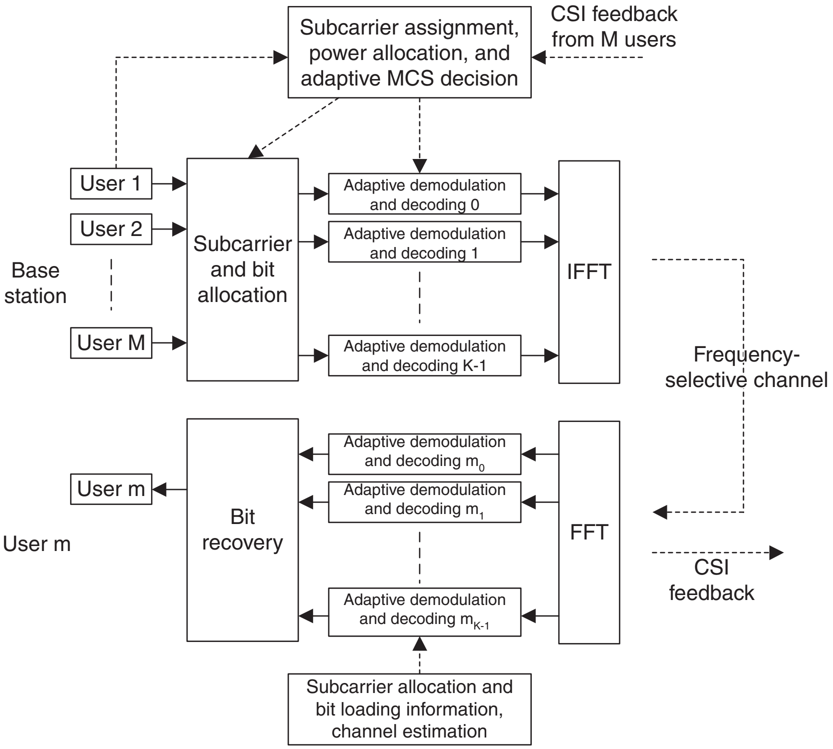

Subcarrier assignment, power allocation, and adaptive MCS decision
CSI feedback from M users
User 1
User 2
Base station
User M
Subcarrier and bit allocation
Adaptive demodulation and decoding 0
Adaptive demodulation and decoding 1
Adaptive demodulation and decoding K-1
IFFT
Frequency-selective channel
User m
User m
Bit recovery
Adaptive demodulation and decoding m_0
Adaptive demodulation and decoding m_1
Adaptive demodulation and decoding m_{K-1}
FFT
CSI feedback
Subcarrier allocation and bit loading information, channel estimation

Figure 4.15 Block diagram of an OFDMA system.

be served by the base station. Based on the channel state information (CSI) report from the M terminals, the base station decides the subcarrier assignment, power allocation, and modulation and coding scheme (MCS) on each subcarrier for each terminal. Then the right number of bits from the queue of each terminal is fed into the subcarriers for adaptive modulation and coding. The modulation order, coding rate and transmitter power will depend on the number of bits to be sent on each subcarrier in each OFDM symbol. Then the complex symbols of all terminals are transformed into the time domain by IFFT for transmission over the frequency-selective channels to all terminals. As each subcarrier has a bandwidth that is significantly smaller than the channel coherence bandwidth, it experiences flat fading. The subcarrier assignment and bit allocation information are sent to all terminals through an independent control channel. After receiving the OFDM symbol, each terminal transforms it back to the frequency domain using FFT. With the information of subcarrier assignment and bit allocation, each terminal is able to demodulate the signal and extract the desired bits. As we can see, additional signal overhead is necessary as all terminals need to feed back CSI and the base station has to inform all terminals of the resource allocation result. However, the overhead is rather small, especially in slow-fading channels, and the resource allocation may be updated once every many OFDM symbols. There are also many other techniques to further reduce the signaling overhead, like bundling of adjacent subcarriers. In the following we focus on the resource allocation aspect of OFDMA systems.

4.9.1 Max throughput scheduling in OFDMA

In the previous sections, we have introduced round-robin, max throughput, PF, max min, and max utility scheduling. All these schedulers can be extended to OFDMA and the corresponding scheduling rules can be easily derived. For example, we first take a look at how max throughput scheduling can be implemented in OFDMA.

The goal of max throughput scheduling is to schedule the terminals such that the total network throughput is maximized. The expected throughput of terminal i in the tth OFDM symbol is

$$R_i[t] = \sum_j W \log_2\left(1 + \frac{P_{ij}[t]\chi_{ij}[t]}{\theta}\right) I(i,j), \tag{4.32}$$

where W is the subcarrier bandwidth and $P_{ij}[t]$ the power allocation on the jth subcarrier of terminal i. $\chi_{ij}[t]$ is the corresponding ratio between the channel gain and the interference plus noise. Therefore, $P_{ij}[t]\chi_{ij}[t]$ is the SINR. $I(i,j)$ is the scheduling indicator and is 1 if the jth subcarrier is assigned to terminal i and 0 otherwise. Each subcarrier is assigned to only one terminal and we have the following constraint,

$$\sum_i I(i,j) = 1, \forall j. \tag{4.33}$$

The overall network throughput is

$$R[t] = \sum_{i=1}^{M} R_i[t] = \sum_i \sum_j W \log_2 \left(1 + \frac{P_{ij}[t]\chi_{ij}[t]}{\theta}\right) I(i,j). \tag{4.34}$$

First, assume there is no adaptive power allocation. In order to maximize the network throughput, subcarrier j should be allocated to the terminal i whose SINR on this subcarrier is the highest among all terminals. That is,

$$I^*(i,j) = \begin{cases} 1 & \frac{P_{ij}[t]\chi_{ij}[t]}{\theta} \geq \frac{P_{mj}[t]\chi_{mj}[t]}{\theta}, \forall m; \\ 0 & \text{otherwise.} \end{cases} \tag{4.35}$$

When adaptive power allocation is used, each subcarrier should be assigned to the terminal with the highest $\chi_{ij}[t]$, as the rate increase by using any amount of power on any subcarrier will be maximized if the subcarrier has the highest $\chi_{ij}[t]$. Therefore,

$$I^*(i,j) = \begin{cases} 1 & \chi_{ij}[t] \geq \chi_{mj}[t], \forall m; \\ 0 & \text{otherwise.} \end{cases} \tag{4.36}$$

The base station still needs to decide how much power should be used on each subcarrier. Given the total transmit power constraint, the problem is to find the power allocation $\mathbf{P}[t] = \{P_{ij}[t]\}$ such that the total network throughput is maximized, i.e.

$$\begin{aligned} \mathbf{P}^*[t] &= \arg_{\mathbf{P[t]}} \sum_i \sum_j W \log_2 \left(1 + \frac{P_{ij}[t]\chi_{ij}[t]}{\theta}\right) I^*(i,j), \\ \text{s.t.} &\sum_{i,j} P_{ij}[t] \leq P_o, \end{aligned} \tag{4.37}$$

where P_o is the total transmit power limit of the base station power amplifier. The objective function in (4.37) is concave in $\mathbf{P}[t]$ and the Lagrangian method can be used to obtain the optimal power allocation. The resulting optimal power allocation is

$$P_{ij}[t] = \left[\lambda - \frac{\theta}{\chi_{ij}[t]}\right]^+ I^*(i,j), \tag{4.38}$$

where $[z]^+ = \max(0,z)$ and λ is determined by the total power limit such that

$$\sum_{i,j} P_{ij}[t] = \sum_{i,j} \left[\lambda - \frac{\theta}{\chi_{ij}[t]}\right]^+ I^*(i,j) = P_o. \tag{4.39}$$

This optimal power allocation is frequently called the water-filling power allocation [J. Jang and K. B. Lee, 2003] and is illustrated in Figure 4.16 when there are two terminals in the network. The x axis is the subcarrier index and y the inverse of channel quality $\frac{\theta}{\chi[t]}$. In the third plot, the terminal with higher channel gain is always scheduled in each subcarrier and the corresponding shadowing part indicates how much power is allocated on the subcarrier for the terminal. As we can see in the figure, more power

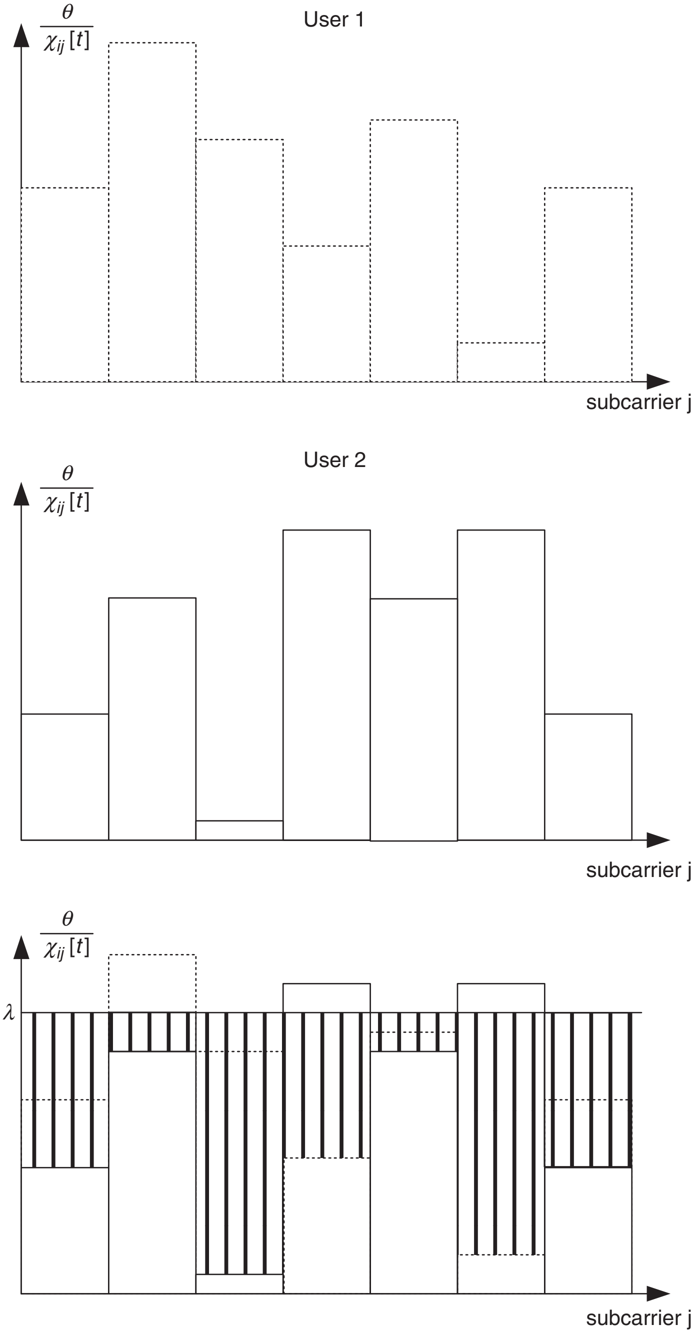

$\frac{\theta}{\chi_{ij}[t]}$
User 1
subcarrier j
$\frac{\theta}{\chi_{ij}[t]}$
User 2
subcarrier j
$\frac{\theta}{\chi_{ij}[t]}$
λ
subcarrier j

Figure 4.16 Water-filling power allocation.

should be allocated to subcarriers with better channel qualities with the water-filling power allocation. Note that the allocation in (4.38) cannot be expressed in closed form and numerical algorithms are necessary to find the optimal λ and correspondingly the power allocation.

Similar to the discussion in Section 4.5, max throughput scheduling for OFDMA also suffers from unfairness and coverage limitations. Terminals closer to the base station have more opportunities for subcarrier assignment and terminals at cell edges may never be served. Therefore fair scheduling algorithms for OFDMA systems have been extensively studied in the literature. For example, the work in [W. Rhee and J.M. Cioffi, 2000] considers throughput maximization with a proportional rate constraint. If we further consider QoS of mobile terminals, the scheduler will be further complicated. For example, when there are terminals with only real-time traffic in the network, the scheduling algorithm can be designed to minimize the total base station power consumption [C. Y. Wong et al., 1999]. The subcarriers, bits loaded on each subcarrier, and power allocation should be jointly designed such that the achieved data rates of all terminals are the same as what they desire so that the total power is minimized. For mixed types of data traffic, utility-based scheduling should be used [G. W. Miao and N. Himayat, 2008]. In general these schedulers have no closed-form expressions. They are usually based on optimization techniques and are much more complicated than the schedulers we have introduced so far in this chapter. Readers interested in the details are referred to [W. Rhee and J.M. Cioffi, 2000; C. Y. Wong et al., 1999; G. W. Miao and N. Himayat, 2008] and the references therein.

Example 4.5: Consider an OFDMA system with two terminals accessing the base station. The channel is divided into four subchannels. On subchannel j, the data rate of terminal 1 is given by $r_{1j} = 5g_{1j}$ kbps, and terminal 2 by $r_{2j} = 3g_{2j}$ kbps, where g_{ij} is the link gain on the subchannel. The link gains for four slots are shown below. Determine the network throughput for the four slots when the following schedulers are applied:

a) round-robin;
b) max throughput.

Channel of terminal 1

Subchannel	Slot 1	Slot 2	Slot 3	Slot 4
1	2.12	2.43	2.52	2.33
2	0.11	0.32	0.90	1.4
3	3.22	2.53	3.01	1.12
4	1.11	1.40	2.00	2.50

Channel of terminal 2

Subchannel	Slot 1	Slot 2	Slot 3	Slot 4
1	1.12	2.43	2.52	4.33
2	13.13	10.36	10.91	11.4
3	5.12	4.23	4.01	5.72
4	12.61	11.43	12.20	12.10

Solution: a) The round-robin scheduler assigns subchannels 1 and 3 to terminal 1 and subchannels 2 and 4 to terminal 2. The average throughputs of the two terminals are

$$\begin{aligned} S_1 = (2.12 + 3.22 + 2.43 + 2.53 + 2.52 + 3.01 + 2.33 \\ + 1.12) \times 5/4 = 24.1 \text{ kbps}; \end{aligned} \tag{4.40}$$

and

$$\begin{aligned} S_2 = (13.13 + 12.61 + 10.36 + 11.43 + 10.91 + \\ 12.20 + 11.4 + 12.10) \times 3/4 = 70.605 \text{ kbps}; \end{aligned} \tag{4.41}$$

The network throughput is $S = S_1 + S_2 = 94.705$`kbps`.

b) The achievable data rates on all subchannels of the two terminals are given below.

Achievable rate of terminal 1

Subchannel	Slot 1	Slot 2	Slot 3	Slot 4
1	10.6	12.15	12.6	11.65
2	0.55	1.6	4.5	7
3	16.1	12.65	15.05	5.6
4	5.55	7	10.00	12.50

Achievable rate of terminal 2

Subchannel	Slot 1	Slot 2	Slot 3	Slot 4
1	3.36	7.29	7.56	12.99
2	39.39	31.08	32.73	34.2
3	15.36	12.69	12.03	17.16
4	37.83	34.29	36.60	36.30

With the max throughput scheduler, the subchannels will be assigned to the terminal with a higher rate and the channel assignment is given as follows:

Channel assignment

Subchannel	Slot 1	Slot 2	Slot 3	Slot 4
1	1	1	1	2
2	2	2	2	2
3	1	2	1	2
4	2	2	2	2

The average throughputs of the two terminals are

$$\begin{aligned} S_1 &= (10.6 + 16.1 + 12.15 + 12.6 + 15.05)/4 \\ &= 16.625 \text{ kbps} \end{aligned} \tag{4.42}$$

and

$$\begin{aligned} S_2 = {} & (39.39 + 37.83 + 31.08 + 12.69 + 34.29 + 32.73 \\ & + 36.6 + 12.99 + 34.2 + 17.16 + 36.30)/4 \\ & = 81.315 \text{ kbps.} \end{aligned} \tag{4.43}$$

The network throughput is $S = S_1 + S_2 = 97.94$ kbps.

Exercises

4.1 Prove that the max throughput scheduler in Section 4.5 also maximizes the total network long-run throughput.

4.2 Derive the PF scheduler in (4.15).

4.3 In a downlink single-cell cellular network, all terminals experience free space line-of-sight channels to the base station. In each time slot, only one terminal can be scheduled and the terminal will send a packet across the whole system bandwidth, 1 MHz. The SINR gap is $\theta = 1$. The transmission power is 1 watt. The noise spectral density is -192.5 dBm/Hz. Five terminals are located at 10, 50, 100, 150 and 250 meters away from the base station. The carrier frequency is 2 GHz. The max throughput scheduler is used.

1) What is the throughput of each terminal?
2) What is the total network throughput?
3) A sixth terminal moves from the cell edge to the cell center and stops for services at the following spots: 200 meters, 100 meters, 20 meters and 5 meters away from the base station. What are the respective throughputs this terminal will receive and the corresponding network throughputs?

4) Draw the individual throughput with respect to the distance to the base station for the sixth terminal as well as the corresponding total network throughput.

4.4 Repeat Exercise 4.3 assuming the round-robin scheduler.

4.5 Assume the same scenario as in Exercise 4.3 and use the PF scheduler. $\tau = 100$. At time t, all terminals have the same long-run throughput of 0.2 Mbps.

1) What is the long-run throughput at time $t + 1$.
2) At time $t + 2$, will another terminal be scheduled? If not, at what time will another terminal be scheduled?

4.6 Derive the PF scheduler for OFDMA using techniques similar to those in Section 4.6.

References

J. Jang and K. B. Lee. 2003 (Feb.). Transmit power adaptation for multiuser OFDM systems. *IEEE J. Sel. Areas Commun.*, 21, 171–178.

N. Jindal, S. Vishwanath and A. Goldsmith. 2004. On the duality of Gaussian multiple-access and broadcast channels. *IEEE Trans. Inf. Theory*, 50(5), 768–783.

B. G. Lee, D. Park and H. Seo. 2009. *Wireless Communications Resource Management*. Singapore: John Wiley & Sons (Asia) Pte Ltd.

G. W. Miao and Z. Niu. 2006. Bandwidth management for mixed unicast and multicast multimedia flows with perception based QoS differentiation. Pages 687–692 of: *Proc. IEEE ICC 2006*.

G. W. Miao and N. Himayat. 2008 (Mar.). Low complexity utility based resource allocation for 802.16 OFDMA systems. Pages 1465–1470 of: *Proc. IEEE WCNC 2008*.

J. Mo and J. Walrand. 2000. Fair end-to-end window-based congestion control. *IEEE/ACM Trans. Networking*, 8(5), 556–567.

W. Rhee and J. M. Cioffi. 2000. Increasing in capacity of multiuser OFDM system using dynamic subchannel allocation. Pages 3866–3875 of: *Proceedings IEEE Vehicular Technology Conference.*

D. Tse and P. Viswanath. 2005. *Fundamentals of Wireless Communication*. Cambridge: Cambridge University Press.

C. Y. Wong, R. S. Cheng, K. B. Lataief and R. D. Murch. 1999. Multiuser OFDM with adaptive subcarrier, bit, and power allocation. *IEEE J. Sel. Areas Commun.*, Oct., 1747–1758.

5 Principles of cellular systems

5.1 Introduction

In this chapter we take a closer look at the interference interaction between different radio communication links that share the same radio spectrum. As a first resource management scheme we consider static allocation when radio links are sharing the available bandwidth via proper orthogonal waveforms. The static allocation scheme is an excellent example to demonstrate most of the features and problems in resource management. It also serves as a reference for most more advanced schemes. Section 5.3 introduces a different resource management scheme to deal with interference in wireless networks. This resource management scheme, random channel allocation, is based on interference averaging by employing proper interference margins.

5.2 Orthogonal multiple access cellular systems

We start with a wireless network having a set of orthogonal waveforms. The design of wireless networks consists of two steps: coverage planning followed by frequency allocation.

5.2.1 Coverage planning

With signal power decreasing with the distance, it is easy to prove that the terminals should establish connections to the (geometrically) closest base station in order to maximize the received SINR. The service area may be partitioned into connection regions surrounding each base station. The connection region of a base station is the geometrical region where the received signal power from that base station is larger than that from any other base station and high enough to meet the quality requirements. Hence, the connection region is always included in the coverage area of the base station. The coverage area of each base station is referred to as *a cell*. The coverage planning problem is to find the required number of base stations to be used within the service area. For the sake of simplicity, assume that the terminals are located on a planar surface with a circular symmetric path loss. Requiring the same transmission quality in all coverage regions clearly means that these regions have to be of uniform shape and size. The most common model used for wireless networks is uniform hexagonal-shaped areas, called

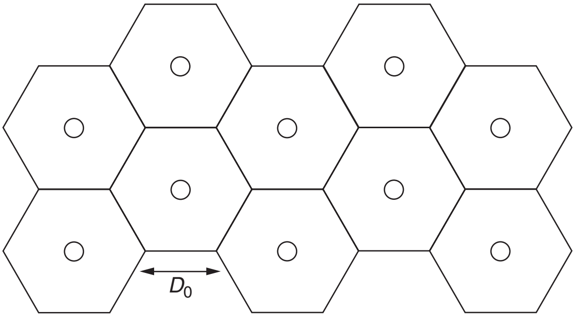


Figure 5.1 Cellular system coverage area layout.

cells. Figure 5.1 shows the geometry of the (hypothetical) coverage regions of such a hexagonal cellular system. A base station with omni-directional antenna is positioned in the middle of each cell.

The area coverage planning is obtained by taking one cell as a reference and then computing the required cell radius such that the signal quality is satisfied over all the cell area. The required received signal-to-noise ratio (SNR) at the border of the cell is ensured if the cell radius satisfies the expression

$$\frac{c_t P_t}{D_0^\alpha N} \geq \gamma_0, \tag{5.1}$$

where D_0 is the radius of the cell, P_t is the transmitted power, N is the additive noise power and c_t is a constant that includes the antenna gains and the path loss constant. Solving for D_0 in (5.1) we get

$$D_0 \leq \left(\frac{c_t P_t}{\gamma_0 N}\right)^{1/\alpha}.$$

Once the radius of the cell is known, the number of cells required to cover the service area can be calculated as

$$\text{Number of cells} = \frac{\text{total area}}{\text{cell area}}.$$

For the particular case of hexagonal cells we get

$$\text{Number of cells} = \frac{\text{total area}}{\frac{3\sqrt{3}D_0^2}{2}}.$$

In the presence of shadow fading, the SNR at the border of the cell will be satisfied only for a certain probability (time availability). Modeling the shadow fading as a log-normal fading process, the outage probability is written as follows:

$$\Pr(\Gamma \leq \gamma_0) = \Pr\left(\frac{c_t P_t m}{r^\alpha N} \leq \gamma_0\right) \leq p_{\text{out}},$$

where m is log-normal distributed with standard deviation σ dB, p_{out} is the required outage probability and r is the distance between the base station transmitter and the terminal. To provide protection against the shadow fading and preserve the same cell radius D_0, an extra fade margin is needed at the transmitter. Denoting by M this extra fade margin and assuming that the terminal is at the border of the cell (worst case), the outage probability becomes

$$\Pr(\Gamma \le \gamma_0) = \Pr\left(\frac{c_t P_t M}{D_0^\alpha N} \le \gamma_0 = \frac{c_t P_t}{D_0^\alpha N}\right) \le p_{\text{out}}.$$

Solving for M we get

$$10\log_{10}(M) = \sigma Q^{-1}(p_{\text{out}}) \text{ dB},$$

where $Q(\cdot)$ is the Q-function.

5.2.2 Static channel allocation

In today's wireless networks the number of simultaneous connections (links) within the service area is larger than the number of orthogonal waveforms (produced by the available bandwidth). An efficient way of managing the radio spectrum and ensuring more radio connections within the service area is to reuse the same frequency within the service area as often as possible. This frequency reuse is possible thanks to the propagation properties of radio waves. In fact, by reusing the same waveforms in different parts of the service area, the mutual interference can be made quite low with proper distance separations. By reusing the waveforms, several connections can be established at the same time and the system capacity can be increased. An apparent paradox is that the higher the propagation loss as a function of distance, the lower the interference level and the more often we can reuse the spectrum. This is illustrated in the following example.

Example 5.1: Consider the situation of Figure 5.2 where two mobile radio connections are established along a highway. The transmitters use the same transmitter power and identical waveforms. The modulation and detection schemes are such that the SINR at the receivers has to be at least 15 dB to achieve a good link quality. Assume that the propagation loss is modeled as a power law distance dependence. Determine the minimum distance D_{12} so that the receiver M_2 can reach its SINR requirement if the propagation constant $\alpha = 4$ and 2, respectively.

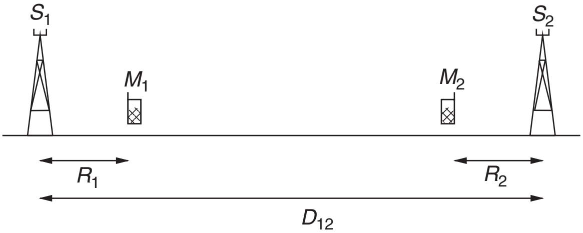


Figure 5.2 Geometry for Example 5.1.

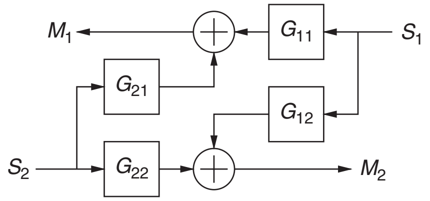


Figure 5.3 System model of the communication system in Example 5.1.

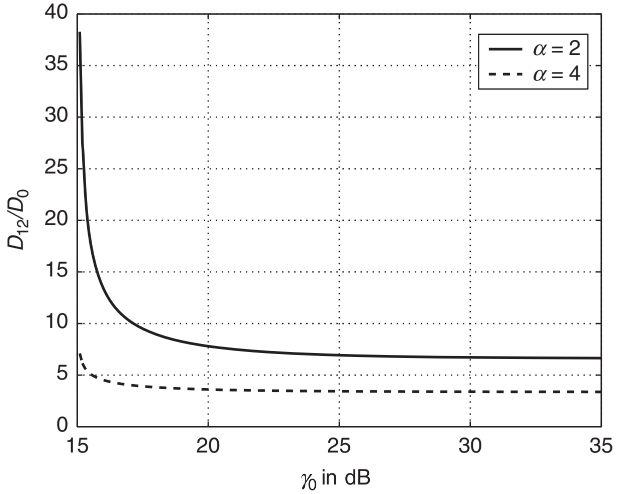


Figure 5.4 Normalized reuse distance as a function of the average SNR at the border of the cell.

Solution: The system can be described schematically by the block diagram of Figure 5.3. The received power may be written as

$$P_r = GP_t = \frac{cP_t}{r^\alpha},$$

where P_t is the transmitter power and r is the distance between the transmitter and the receiver. Using this expression, the received SINR at receiver M_2, denoted by Γ_2, can be written as follows:

$$\Gamma_2 = \frac{\frac{cP_2}{R_2^\alpha}}{\frac{cP_1}{(D_{12}-R_2)^\alpha} + N} = \frac{1}{\frac{P_1}{P_2}\left(\frac{R_2}{D_{12}-R_2}\right)^\alpha + \frac{1}{\gamma_2}} \geq \gamma_t, \tag{5.2}$$

where P_i is the transmitter power of transmitter S_i, γ_t is the required SINR threshold for good signal quality, and $\gamma_2 = \frac{c_t P_2}{R_2^\alpha N}$. If the transmitter powers are equal, $P_1 = P_2 = P_t$, the received SIR reduces to

$$\Gamma_2 = \frac{1}{\left(\frac{R_2}{D_{12}-R_2}\right)^\alpha + \frac{1}{\gamma_0}} \geq \gamma_t,$$

where $\gamma_0 = \frac{c_t P_t}{D_0^\alpha N}$ with $D_0 = R_1 = R_2$. For an SINR threshold of 15 dB, we obtain a separation distance given by

$$D_{12} \geq D_0 \left[1 + \frac{1}{\left(\frac{1}{\gamma_t} - \frac{1}{\gamma_0}\right)^{1/\alpha}} \right].$$

Figure 5.4 illustrates the normalized reuse distance as a function of the average received SNR at the border of the cell. It is observed that by increasing γ_0 the reuse distance reaches its minimum and the system becomes interference limited (N can be neglected in the presence of interference).

Note that the received SINR is a deterministic quantity and the SIR is always achieved as soon as the distance relation above is satisfied. In the presence of shadow fading, the received SINR becomes a random variable and the reuse distance is determined for a given outage probability.

The minimum physical distance between two transmitters using the same waveform, required to achieve a certain link quality, is usually termed *the reuse distance*. The inverse of this distance is a rough measure of how many radio links are able to reuse the available bandwidth per unit length. As noted in the example, the SINR increases rapidly with increasing distance D_{12}. A very important observation is that the SINR does not depend on the absolute distances, but only on the ratio between D_{12} and R_2. This means that the system is scalable, whereby we can change the distance scales without affecting the SIR. However, the scaling process cannot be driven to extremes since, at very short distances, the propagation conditions approach (line-of-sight) free-space conditions ($\alpha = 2$). Example 5.1 also illustrates that the reuse distance decreases with increasing propagation exponent α. It is also observed that the reuse distance depends on the minimum required SINR. This quantity depends on the modulation and detection schemes employed. Waveforms and detectors more robust to interference will allow for frequent (dense) reuse.

Example 5.1 clearly shows that the SINR constraints can be satisfied by using short communication distances and by keeping the other interfering users far away. The problem can be generalized to a large number of base stations that are dispersed in a regular fashion within the service area. With signal power decreasing with the distance, it is easy to prove that the terminals should establish connections to the closest base station in order to maximize the received SINR. With N base stations using the same waveform within the service area, the received SINR at a given mobile terminal within cell k can be written as follows:

$$\Gamma_k = \frac{\frac{c_t P_k}{r^\alpha}}{\sum\limits_{i=1, i\neq k}^{N} \frac{cP_i}{D_{i,k}^\alpha} + N},$$

where P_i is the transmitter power of base station i and $D_{i,k}$ is the distance between base station i and mobile terminal k. With a proper choice of the reuse distances $D_{i,k}$, good link quality can be achieved for all active links within the different cells that are using the same frequencies.

One of the objectives in wireless network design is to achieve a system with the highest capacity possible for a given available bandwidth: a system that allows for the transmission of most information or that will support as many simultaneous connections as possible.

The classical resource allocation scheme that is found in all early mobile telephone systems [V. H. MacDonald, 1979; D. C. Cox, 1982] is the static resource allocation scheme or fixed channel allocation (FCA). In this scheme, each access port is assigned a certain fixed number of channels. If the expected number of active terminals is the same in all cells of the network, the number of channels assigned to each access port should also be the same to provide the same level of service in all parts of the system. This allocation is achieved by dividing the C available channels into K groups of (approximately) equal size where each access port (cell) is assigned a group of

$$\eta = \left\lfloor \frac{C}{K} \right\rfloor \text{ channels/cell,} \tag{5.3}$$

where $\lfloor x \rfloor$ denotes the integer part of x. The access port has the right to use these channels freely to communicate with its terminals, but cannot use any channel from another group. The group size η is a raw measure of the capacity of the system since it indicates the maximum number of simultaneous connections that can be supported in each cell. For instance having a total number of available channels $C = \{s_0(t), s_1(t), \ldots, s_{C-1}(t)\}$ we get K disjoint channel groups $C_i = \left\{ s_{i\frac{C}{K}}, s_{1+i\frac{C}{K}}, \ldots, s_{(i+1)\frac{C}{K}-1} \right\}$, $i = 0, 1, \ldots, K-1$.

As the received power is monotonically decreasing with distance, the interference that will be caused by reusing the channels over and over again in the system will be dependent on how far away the same channel is reused. To maintain a sufficiently high SINR in the base station–terminal connections, the channels in a group cannot be reused in a cell that is too close to the first cell. We characterize a channel assignment by its *(minimum) reuse distance* D, which is the smallest distance between two cells using the same channel. Clearly, if many channel groups are used (i.e. large K), the reuse distance will be large which gives a relatively high SINR. The penalty paid for such a procedure is, as (5.3) illustrates, that the number of channels that each access port has at its disposal becomes small. The capacity of the system will then be low. On the other hand, if a low SINR can be allowed, D can be made small. In this case only a few channel groups are necessary and the raw capacity η will be higher. Evidently, there is a tradeoff between transmission quality and transmission capacity.

Assume that the K disjoint waveform groups are numbered $1, 2, \ldots, K$. Each cell is thus assigned one of these numbers. A good allocation procedure should provide a reuse symmetry within the service area and maximize the minimum distance between two cells with the same channels (i.e. maximizing D for a given K). Alternatively, for a given minimum reuse distance D find the minimum number of required channel groups K. The allocation process then consists of forming a symmetric cluster of the channel groups and repeating the cluster over the service area until the service area is filled. Figure 5.5 shows some clusters of different sizes for hexagonal cells.

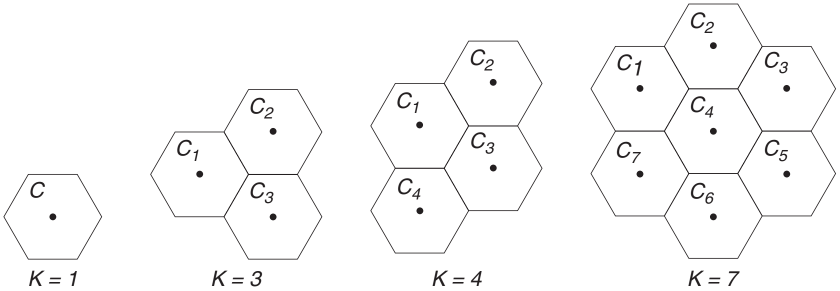


Figure 5.5 Examples of cluster sizes for $K = 1, 3, 4$ and 7 for hexagonal cells.

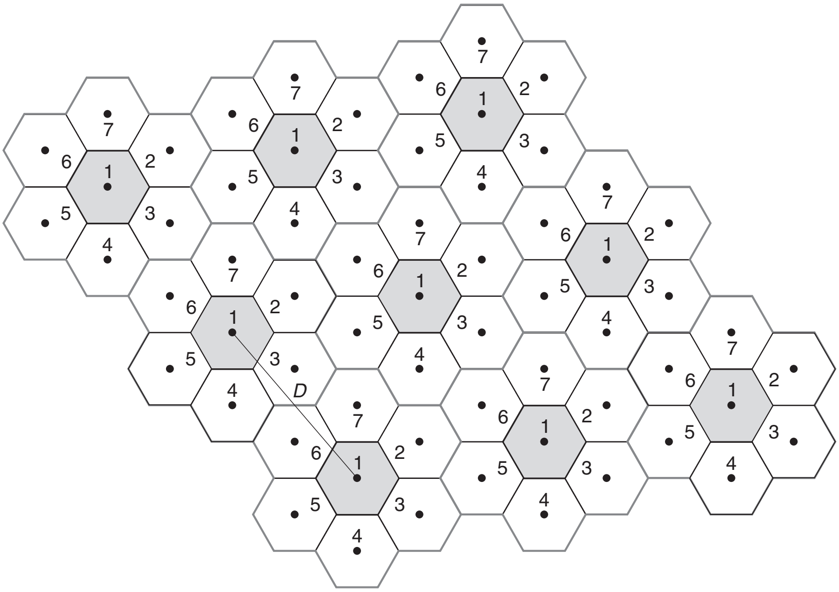


Figure 5.6 Examples of symmetric hexagonal cell plans.

The values of K used in Figure 5.5 will give a *fully symmetric cell plan*. Figure 5.6 shows the symmetric hexagonal cell plan when $K = 7$. These cell plans are characterized by the property that the patterns formed by the cells having the same label are identical (just shifted) for all labels. Further, in such a cell plan, each cell has six nearest neighbors, all at the same (minimum reuse) distance D. For hexagonal fully symmetric cell plans one can show that the following relationship between K and D holds:

$$\Delta = \frac{D}{D_0} = \sqrt{3K}, \tag{5.4}$$

where D_0 denotes the radius of the hexagonal cell. The quantity Δ is the *normalized reuse distance*. It is possible to show that there exist fully symmetric cell plans for all (integer) K that can be written in the form

$$K = (i + j)^2 - ij, \quad i, j = 0, 1, 2, 3, \ldots \tag{5.5}$$

Possible values for K are thus $K = 1, 3, 4, 7, 9, 12, 13, \ldots$, where $K = 1$ corresponds to the trivial case where all channels are used in all cells. For a more detailed discussion and a mathematical analysis of such cell plans, the reader is referred to [R. A. Leese, 1997]. Note that these values of K are valid for hexagonal cell shapes only. Other cell shapes such as rectangular or triangular will have different valid values for the cluster size K. The principle of finding K is, as mentioned earlier, to ensure symmetry in the cell plan.

5.2.3 Capacity analysis—guaranteed service

In the following we make some highly simplifying assumptions in a first attempt to assess the capacity of a guaranteed service system (e.g. a mobile telephony system) based on static resource allocation. The analysis is similar to the one found in [W. C. Y. Lee, 1993].

The capacity of a wireless network is measured as the average number of simultaneous radio links supported by the system and is given by

$$\eta = \frac{C}{K} \text{ users/cell.}$$

Taking the covered area into account we define the area capacity as follows:

$$\eta_A = \frac{C}{KA_{\text{cell}}} \text{ users/unit area,}$$

where A_{cell} is the cell area. Hence, computing the capacity of a hexagonal cellular system reduces to computing the cluster size K (or equivalently the reuse distance D).

Let us assume that our mobile telephony system uses a symmetrical hexagonal cell plan and a modulation scheme requiring a minimum SINR γ_t to achieve acceptable link performance. The transmitters use constant transmit power P and without loss of generality we will study the downlink of the system. Let us therefore assess the SINR at some given terminal in a cell using a specific channel. To guarantee the minimum SINR for this terminal, we will look at the worst-case situation where the channel is used in all these co-channel cells. Further, it is easy to convince oneself that the lowest SINR in the center cell is found when a terminal is on the cell boundary, in one of the corners of the hexagon. The SINR at the mobile terminal (see Figure 5.7) can be expressed as

$$\Gamma = \frac{\frac{cP}{D_0^\alpha}}{\sum_k \frac{cP}{D_k^\alpha} + N} = \frac{1}{\sum_k \left(\frac{D_0}{D_k}\right)^\alpha + \frac{1}{\gamma_0}} \tag{5.6}$$

where N is the noise power, c is an arbitrary propagation- and antenna-related constant, and where we have introduced γ_0 with

$$\gamma_0 = \frac{cP}{ND_0^\alpha}$$

which is the SNR at the terminal, i.e. at the cell boundary.

We can now distinguish between some special cases:

1. *The noise/range limited case:* The interference is much smaller than the thermal noise. One could say that in this case the APs are very sparsely placed, which either is due to low traffic demand and/or the system has coverage problems. This is a typical situation when systems are first deployed and in rural areas. Capacity is not the issue in these systems, coverage is.
2. *The interference limited case:* The interference is much stronger than the thermal noise. There is plenty of received signal power and there are no coverage problems. This is the typical situation in a mature, highly loaded system in urban areas. Capacity is the key issue in these systems, coverage is not.

In a symmetric hexagonal cell plan each cell has exactly six co-channel neighbors (using the same channel group) at a distance D. Furthermore, there are six additional co-channel cells at a distance $\sqrt{3}D$, six at a distance $\sqrt{4}D$, six at a distance $\sqrt{7}D, \ldots,$ six at a distance $\sqrt{K}D$, and so on, for all values of K given by the expression (5.5) (see Figure 5.7). We further assume that all interferers within the same interference tier are at the same distance. It can be noted that some of the interferers will actually be closer than the nominal distance, some will be farther away. Unless very small reuse distances are used, this approximation will thus not introduce very large errors. However, even with this approximation we are not able to obtain a closed form expression for the received SINR at the mobile terminal. Another approximation that allows us to obtain a closed

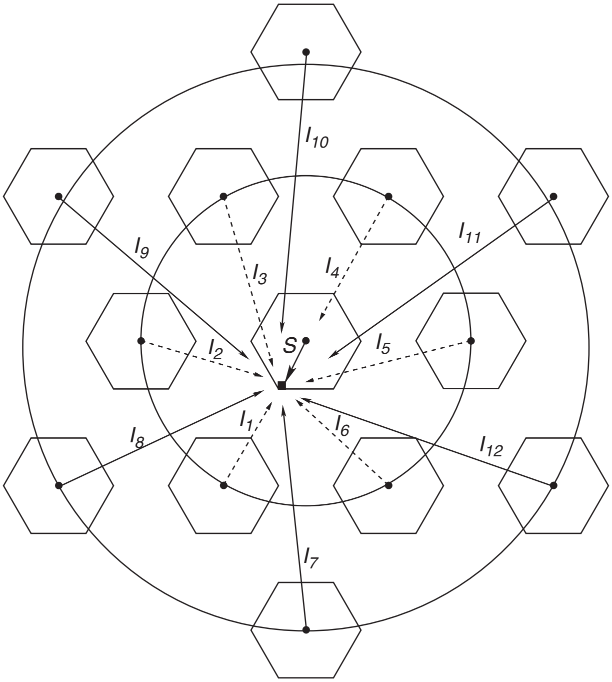


Figure 5.7 Co-channel cell tiers in the downlink of hexagonal cellular systems.

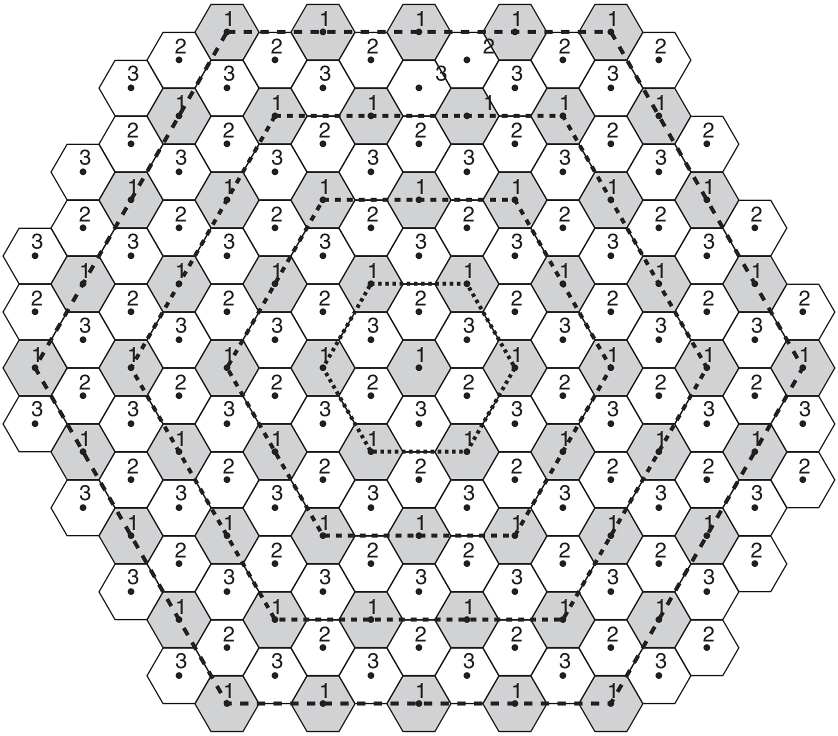


Figure 5.8 Co-channel interference tiers illustration when $K = 3$.

form expression is to view the co-channel interferers as illustrated in Figure 5.8, where one can assume that there are 6 interferers at a distance D from the cell of interest, 12 interferers at a distance $2D$, 18 interferers at a distance $3D$, and so on. With that, the received SINR can be written as

$$\begin{aligned}\Gamma_{\mathrm{DL}} &\approx \frac{1}{\frac{6}{\left(\sqrt{3K}\right)^{\alpha}} + \frac{12}{\left(2\sqrt{3K}\right)^{\alpha}} + \frac{18}{\left(3\sqrt{3K}\right)^{\alpha}} + \cdots + \frac{1}{\gamma_0}} \\ &= \frac{1}{6\sum_{n=1}^{+\infty} \frac{n}{(n\sqrt{3K})^{\alpha}} + \frac{1}{\gamma_0}} \\ &= \frac{1}{\frac{6}{(3K)^{\alpha/2}}\zeta(\alpha-1) + \frac{1}{\gamma_0}},\end{aligned} \tag{5.7}$$

where $\zeta(\cdot)$ is the Riemann zeta function, which has the values

$$\zeta(\alpha-1) = \sum_{n=1}^{+\infty} \frac{1}{n^{\alpha-1}} = \begin{cases} +\infty, & \alpha = 2 \\ \frac{\pi^2}{6} = 1.6449, & \alpha = 3 \\ 1.2021, & \alpha = 4 \\ \frac{\pi^4}{90} = 1.0823, & \alpha = 5 \end{cases}.$$

For large values of α ($\alpha \geq 4$), we notice that the function $\zeta(\alpha)$ converges to 1, meaning that the co-channel interference is dominated by the six closest interferers.

Introducing the SINR requirement yields:

$$\frac{1}{\frac{6}{(3K)^{\alpha/2}}\zeta(\alpha-1)+\frac{1}{\gamma_0}} \geq \gamma_t \quad \rightsquigarrow \quad K \geq \frac{1}{3}\left(\frac{6\zeta(\alpha-1)}{\frac{1}{\gamma_t}-\frac{1}{\gamma_0}}\right)^{2/\alpha} \tag{5.8}$$

We thus have to pick the first K in the sequence (5.5) that fulfils the condition (5.8).

A similar analysis can be made in the uplink; the uplink received SINR at the base station is approximated as

$$\begin{aligned}\Gamma_{\mathrm{UL}} &\approx \frac{1}{\frac{6}{(\sqrt{3K}-1)^{\alpha}}+\frac{12}{\left(2\sqrt{3K}-1\right)^{\alpha}}+\frac{18}{\left(3\sqrt{3K}-1\right)^{\alpha}}+\cdots+\frac{1}{\gamma_0}} \\ &= \frac{1}{6\sum_{n=1}^{+\infty}\frac{1}{(n\sqrt{3K}-1)^{\alpha}}+\frac{1}{\gamma_0}}.\end{aligned}$$

Assuming that the total experienced interference is dominated by the six closest interferers, the cluster size can be bounded below as

$$K \geq \frac{1}{3}\left(1+\left[\frac{6}{\frac{1}{\gamma_t}-\frac{1}{\gamma_0}}\right]^{1/\alpha}\right)^2. \tag{5.9}$$

Example 5.2: A mobile telephone system with 120 channels uses a modulation scheme requiring a minimum SIR, γ_t, of at least 15 dB to achieve acceptable link performance. Assume that the propagation loss is only distance-dependent and increases with the fourth power of the distance. The system can be assumed to be interference limited. At most how many channels per cell η can be offered by the system?

Solution: Assuming the uplink case and considering the six closest interferers, the SINR is given by

$$\Gamma = \frac{1}{\frac{6}{(\sqrt{3K}-1)^4}} \geq 31.6.$$

Using (5.9) we get

$$K \geq \frac{1}{3}\left[(6\times 31.6)^{1/4}+1\right]^2 = 7.4.$$

Checking the condition in (5.5) we see that there exists a symmetrical cell plan for $K = 1,3,4,7$ and 9. $K = 7$ will obviously not provide a large enough reuse distance, forcing the use of $K = 9$.

The quantity η now becomes

$$\eta = \left\lfloor \frac{C}{K} \right\rfloor = \left\lfloor \frac{120}{9} \right\rfloor = 13 \text{ channels/cell}.$$

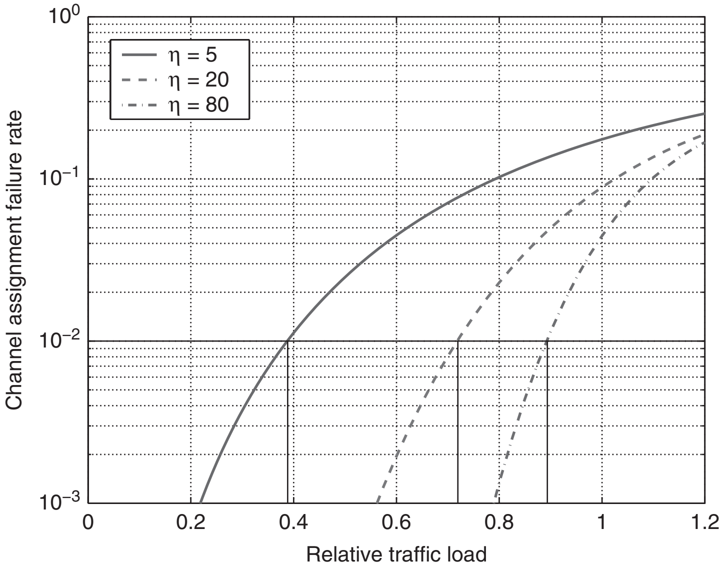


Figure 5.9 Channel assignment failure rate as a function of the relative traffic load ϖ_η for $\eta = 80, 20$ and 5 channels/cell.

5.2.4 Traffic-based capacity analysis

The number of available channels in every access port is a rough but useful capacity measure. It describes the capability to serve users from an operator perspective, but it does not reflect user satisfaction. In general, the number of active users within the cell is not constant but changes with time. In cellular networks, calls are usually modeled as a Poisson process with an expected value of λ calls. Denoting by M_c the random number of mobile terminals in one particular cell at a given observation time interval, we have

$$\Pr(M_c = k) = \frac{(\omega A_c)^k}{k!} e^{-\omega A_c},$$

where ω is the number of calls per unit area, A_c is the cell area, and $\lambda = E\{M_c\} = \omega A_c$ calls.

A cellular system having η channels per cell will experience failure in assigning channels to users when the number of calls exceeds the number of available channels. The assignment failures within the cell can be computed as follows:

$$Z = \max(0, M_c - \eta).$$

The assignment failure rate can then be found as

$$\begin{aligned} \nu_p &= \frac{E\{Z\}}{E\{M_c\}} = \frac{E\{\max(0, M_c - \eta)\}}{\omega A_c} \\ &= \sum_{k=\eta}^{+\infty} (k - \eta) \frac{(\omega A_c)^{k-1}}{k!} e^{-\omega A_c}. \end{aligned} \tag{5.10}$$

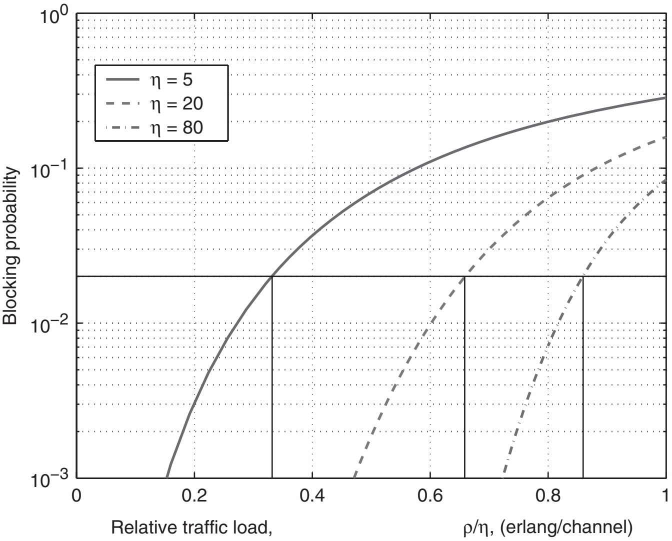


Figure 5.10 Blocking probability as a function of the relative traffic intensity per channel and cell.

We define the relative traffic load ϖ_η as

$$\varpi_\eta = \frac{\omega A_c}{\eta} = \frac{\omega A_c}{C}K.$$

An alternative measure of the traffic load is the expected number of terminals per cell per total number of channels in the entire system, C,

$$\varpi_c = \frac{\omega A_c}{C} = \frac{\varpi_\eta}{K}. \tag{5.11}$$

As observed from (5.10), the assignment failure rate is an increasing function in ω and decreasing in η. Less obvious is the fact that the assignment failure rate is decreasing in η when we increase ω by the same amount, i.e. keeping ϖ_η constant. That is, a system with many channels is more efficient than a system with few channels: a well-known result from traffic theory (the bigger the better; trunking gain). This result is illustrated by the following example.

Example 5.3: A cellular telephone system designed for $K = 9$ channel groups has a cell radius of 1 km. What is the capacity of the system (measured in calls/km^2) if we allow an assignment failure rate of at most 1% and the system has $C = 720, 180$ or 45 channel pairs at its disposal?

Solution: We first determine the area of the cell as

$$A_c = \frac{3\sqrt{3}}{2}1^2 \approx 2.6 \text{ km}^2.$$

The number of available channel pairs/cell is given by ($K = 9$)

$$\eta = \left[\frac{C}{K}\right] = 80, 20 \text{ and } 5 \text{ channel pairs/cell.}$$

From Figure 5.9 we find:

$$\varpi\eta = \frac{\omega A_c}{\eta} = 0.89, 0.72 \; resp. \; 0.39$$

for $\eta = 80, 20$ and 5 respectively at a 1% assignment failure rate. Combining the above obtained results we get a capacity of

$$\omega = \frac{\eta \varpi_\eta}{A_c} = 27, 5.5 \; and \; 0.75 \; calls/km^2.$$

Note that the number of channels/cell is increased by a factor of 4 between the different cases. The capacity increases by a factor of 7 and 5 respectively due to the trunking gain.

Notable in the example is that the capacity is primarily dependent on η, the number of channels per cell and the cell area, A_c. Rewriting the last expression in the example yields

$$\omega = \frac{\eta \varpi_\eta}{A_c} = \frac{C \varpi_\eta}{K A_c}. \tag{5.12}$$

Using the result in (5.8), the following approximation can be made:

$$K = \frac{\Delta^2}{3} \approx c(\alpha) \gamma_t^{2/\alpha},$$

where $c(\alpha)$ is a constant and γ_t is the minimum required SINR. Inserting this result into (5.12), the following approximate result is obtained:

$$\omega \approx c'(\alpha) \frac{C \varpi_\eta}{\gamma_t^{2/\alpha} A_c}, \tag{5.13}$$

or, in dB,

$$\begin{aligned} 10\log_{10}(\omega) \approx 10\log_{10}\left(c'(\alpha)\right) + 10\log_{10}(C) + 10\log_{10}(\varpi_\eta) \\ -\frac{20}{\alpha}\log_{10}(\gamma_t) - 10\log_{10}(A_c), \end{aligned}$$

where $c'(\alpha)$ is a constant.

As we can see from the expression in (5.13), there are in principle three ways to increase the capacity of a wireless communication system:

- *Increasing C:* Fairly obvious. More spectrum resources may be hard to come by and could be difficult.
- *Decreasing γ_t:* The required threshold can be reduced by employing more interference-resistant modulation, detection and channel coding schemes. Since this results in a decreased K, the gain is twofold since the factor ϖ_η is also decreased.

- *Decreasing A_c:* There is in fact no real limit to how large the capacity may become when decreasing the size of the cells. Very small A_c will improve propagation (i.e. to line-of-sight conditions $\alpha = 2$) which will reduce the constant c' somewhat. The penalty is increasing the number of cells (base stations) per unit area. Since the number of base stations is inversely proportional to A_c, the capacity will be roughly proportional to the total number of base stations in the system!

Another observation is that poorer propagation conditions, i.e. an increasing α, in fact improves the capacity of the system.

Systems using static resource allocation are also quite easy to analyze using the traditional traffic model for telephony. If the assumption that the mobility of the terminals is rather limited (i.e. the terminals generally stay within a single cell for the duration of each call) is made, the call handling in each cell can be modeled as independent M/M/η blocking systems. This type of queuing system is characterized by calls arriving according to a Poisson process (independent inter-arrival times with exponential distribution). The arriving calls are served by servers (the η available channels per cell). Calls are blocked and disappear if all channels are found to be busy. The classical performance measure here is the *blocking probability*, i.e. the probability that a newly arriving call finds all channels busy and is denied service. The blocking probability is given by the well-known Erlang-B formula

$$E_\eta = \frac{\frac{\rho^\eta}{\eta!}}{\sum_{k=0}^{\eta} \frac{\rho^k}{k!}}, \tag{5.14}$$

where ρ is the traffic load. We define the relative traffic load as

$$\rho_c = \frac{\rho}{\eta}.$$

Similar to telephone channels, the capacity can also be defined as follows:

$$\rho_{\max} = \max\left\{\rho \,\middle|\, E_\eta < p_0\right\} \text{ erlang},$$

where p_0 is the required blocking probability. The area capacity can be obtained as

$$\rho_A = \frac{\rho_{\max}}{A_c} \text{ erlang/unit area}, \tag{5.15}$$

where A_c is the area of the cell.

The question is how different the obtained results are when using this blocking system principle in comparison with the assignment failure rate principle discussed earlier. This is illustrated in the following example.

Example 5.4: A mobile telephony system with slowly moving terminals uses a cell plan with $K = 9$ channel groups and a cell radius of 1 km. What is the capacity of the system (measured in erlang/km^2 at a maximum blocking probability of 2%) if $C = 720, 180$ or 45 channel pairs are available?

Solution: The cell is as obtained in Example 5.3 and the number of available channels per cell is $\eta = 80, 20$ and 5. Using (5.14), the blocking probability for different values of η is illustrated in Figure 5.10 as a function of the relative traffic load. With a blocking probability of 2% we get the following relative traffic loads:

$$\frac{\rho}{\eta} = 0.87, 0.67 \text{ and } 0.33.$$

Solving for the area capacity we get

$$\rho_A = \frac{\rho}{A_c} = \left(\frac{\rho}{\eta}\right)\frac{\eta}{A_c} \approx 27,5 \text{ and } 0.65 \text{ erlang}/km^2,$$

which are quite similar to the numerical results obtained in Example 5.3.

5.2.5 Best-effort data services

Mobile data networks provide variable data rates to the users. Hence, the number of active users within the cell will not affect the capacity of these systems but rather the peak and average data rates. We will try here to relate the cell capacity of a mobile data system with an adaptive data rate to the spectrum reuse properties. Then we will find a relationship between the cost and peak rates in the system in order to get some guidance on how future systems should be deployed in a cost-efficient way. Without loss of generality, we study the downlink in a single circular cell where the expected received power decays with the α-power of the distance to the base station at the center of the cell. We assume that the system is channelized as before. However, the size of the channels may vary. For a hexagonal cellular system with a cluster size K, we have an available bandwidth per cell given by

$$W = \frac{W_s}{K},$$

where W_s is the total bandwidth allocated to the system and K is the cluster size. We assume an ideal transmission system operating at the Shannon rate, adapting to the instantaneous signal-to-interference ratio with a peak data rate $R_{\max}$. We assume that the system is employing a TDMA access scheme. The received SINR for a given active user within the cell can be written as

$$\Gamma_{\mathrm{DL}}(D) = \frac{\frac{cP_t}{D^\alpha}}{\sum_k \frac{cP_t}{D_k^\alpha} + N}, \tag{5.16}$$

where D is the distance between the base station and the mobile terminal.

Defining the normalized distance $d = D/D_0$ with $0 \le d \le 1$, the received SINR can be rewritten as

$$\Gamma_{\mathrm{DL}}(d) = \frac{1}{d^\alpha}\frac{\frac{cP_t}{D_0^\alpha}}{\sum_k \frac{cP_t}{D_k^\alpha} + N} = \frac{\Gamma_{\mathrm{DL}}(D_0)}{d^\alpha}. \tag{5.17}$$

where $\Gamma_{\mathrm{DL}}(D_0)$ is the received SINR at the border of the cell as defined in (5.7). Note that $\Gamma(D_0)$ is the SINR at the cell boundary, which only depends on the cellular design (i.e. the reuse factor K) and the boundary SNR γ_0.

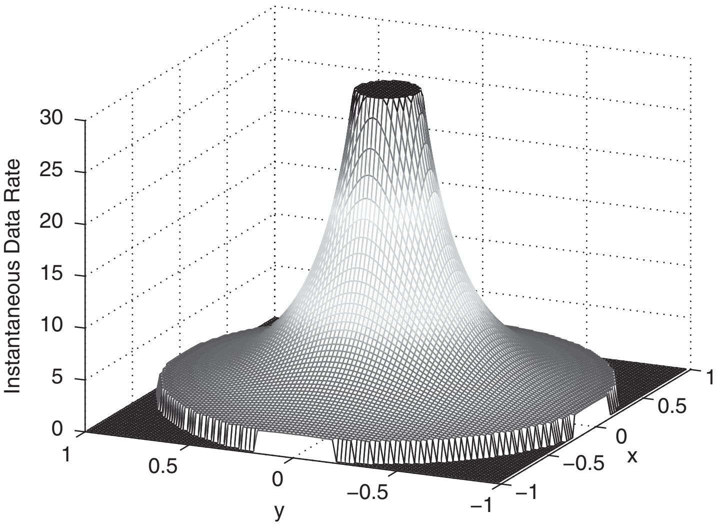


Figure 5.11 Data rates in a cellular (mobile) data systems.

The data rate R of the active user within the cell can then be written as

$$\begin{aligned} R(d) &= \min\left\{R_{\max}, c_w W \log_2\left(1+\Gamma_{\mathrm{DL}}(d)\right)\right\} \\ &= \min\left\{R_{\max}, c_w W \log_2\left(1+\frac{\Gamma(D_0)}{d^\alpha}\right)\right\}, \end{aligned} \tag{5.18}$$

where c_w with $0 \leq c_w \leq 1$ is a constant that accounts for the gap between the capacity upper bound (Shannon capacity) and the actual channel capacity.

Figure 5.11 illustrates the behavior of that data rate as a function of the position in the cell. It is observed that the data rate at the center of the cell is limited by the *peak data rate* $R_{\max}$. As we approach the cell boundary, the data rate drops and at the cell edge ($d = 1$) we have the *edge data rate* (minimum data rate)

$$\begin{aligned} R_{\min} &= \min\left\{R_{\max}, c_w W \log_2\left(1+\Gamma(D_0)\right)\right\} \\ &= c_w W \log_2\left(1+\Gamma(D_0)\right), \end{aligned} \tag{5.19}$$

of course as long as $R_{\max} > R_{\min}$. The edge data rate is thus also defined by the cellular design and the boundary SNR. We may also express the date rate in (5.18) in terms of the minimum data rate as

$$R(d) = \min\left\{R_{\max}, cW \log_2\left(1+\frac{2^{R_{\min}/cW}-1}{d^\alpha}\right)\right\}. \tag{5.20}$$

Let us assume, as in Chapter 2, that the users are uniformly (two-dimensional Poisson process) distributed over a circular approximation of the cell. The expected data rate at

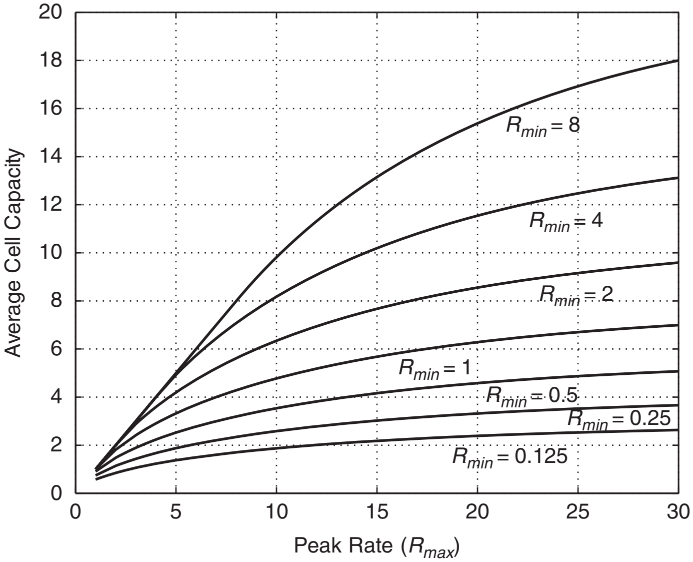


Figure 5.12 Cell capacity as a function of peak data rate R_{max}. All rates are in Mbps with $W = 5$ MHz and $c = 1$.

some randomly picked user will now be

$$\begin{aligned}\bar{R} = E[R] &= \int_0^1 R(x)2x dx \\ &= \int_0^1 2x \min\left\{R_{\text{max}}, cW\log_2\left(1 + \frac{\Gamma(D_0)}{x^\alpha}\right)\right\} dx \\ &= \int_0^1 2x \min\left\{R_{\text{max}}, c_w W\log_2\left(1 + \frac{2^{R_{\text{min}}/c_w W} - 1}{x^\alpha}\right)\right\} dx.\end{aligned}$$

This is in fact the average data rate a user would experience if every user were given an equal amount of resources (e.g. time slots in a round-robin scheduler). We will therefore call it *the cell capacity*. The integral on the right-hand side has, unfortunately, to be numerically calculated. Figure 5.12 shows some numerical examples. What we can see in the figure is that the average rate is to a large extent dominated by the edge rate R_{min}. If R_{min} is low, it really doesn't matter if we have an advanced system that can provide very high peak rates. The following example illustrates a practical problem where data rates are not adjusted in a continuous fashion but rather in discrete steps.

Example 5.5: A mobile data system with $K = 3$ reuse and a total of 15 MHz system bandwidth uses an adaptive modulation scheme operating at 50% of the rate of the Shannon limit ($c_w = 0.5$) with a discrete set of rates $R = 1, 2, 4, 8$ and 16 Mbps and the SNR at the boundary is $\gamma_0 = 6$ dB. The propagation path loss exponent is $\alpha = 4$. Estimate the cell capacity per channel!

Solution: First we compute the received SINR by combining (5.6) and (5.8):

$$\Gamma(D_0) = \frac{1}{\sum_k \left(\frac{D_0}{D_k}\right)^\alpha + \frac{1}{\gamma_0}} \approx \frac{1}{\frac{6\zeta(\alpha-1)}{(3K)^{\alpha/2}} + \frac{1}{\gamma_0}} = \frac{1}{\frac{6\times 1.2021}{(9)^{4/2}} + \frac{1}{4}} \approx 3.$$

As we can see the SINR is dominated by the noise. The channel bandwidth is given by

$$W = \frac{W_s}{K} = \frac{15}{3} = 5 \text{ MHz}.$$

The highest edge data rate of the system is obtained from (5.19) and is

$$R_{\min} = c_w W \log_2\left(1 + \Gamma(D_0)\right) = 2.5\log_2(1+3) = 5 \text{ Mbps}.$$

This means that the highest data rate from our discrete set that can be sustained at the cell edge is 4 Mbps. As we move to the center of the cell the SINR increases and at some distance d_8 we will be able to sustain the next data rate in our sequence, 8 Mbps. From (5.18) we have

$$R(d_8) = 0.5 \times 5 \times \log_2\left(1 + \frac{3}{d_8^\alpha}\right) = 8.$$

Solving for d_8 we get

$$d_8 = \left(\frac{3}{2^{\frac{8}{2.5}} - 1}\right)^{1/4} = 0.78.$$

In the same way we calculate $d_{16} \approx 0.43$. Again approximating the cell with a circle, and remembering that the proportion of users in a certain area is proportional to that area, we now calculate the average data rate:

$$\begin{aligned}\bar{R} = E[R] = \int_0^1 R(x) 2x dx &= 16 d_{16}^2 + 8 \times \left(d_8^2 - d_{16}^2\right) + 4\left(1 - d_8^2\right)\\ &= 16 \times 0.43^2 + 8 \times \left(0.78^2 - 0.43^2\right) + 4\left(1 - 0.78^2\right)\\ &\approx 7.94 \text{ Mbps}.\end{aligned}$$

Note that the peak data rate is usable only below 43% of the cell radius, corresponding to a fraction of 19% ($= 0.43^2$) of the users! Omitting the highest data rate (i.e. if $R_{\max} = 8$), the average data rate is decreased by less than 1 Mbps. This a quite common situation for noise-limited systems.

Special case: If we look at the noise-limited case as we let W and $R_{\max}$ approach infinity we can in fact compute an exact expression for (5.21). We have

$$\lim_{W, R_{\max} \to +\infty} R(d) = \lim_{W, R_{\max} \to +\infty} cW \log_2\left(1 + \frac{P}{d^\alpha N_0 W}\right) = c\frac{P}{d^\alpha N_0}.$$

The average data rate becomes

$$\begin{aligned}\bar{R}_\infty = E[R] = \int_0^{\Delta_0} R(x)2x dx = \int_0^{d_1} 2x R_{\max} dx + \int_{d_1}^{1} 2x \frac{c_1}{d^\alpha} dx \\ = (2\alpha - 3) R_{\max} \left(\frac{R_{\min}}{R_{\max}} \right)^{2/\alpha} - 2(\alpha - 2) R_{\min}. \end{aligned} \tag{5.21}$$

Note that the average data rate in this case is not bounded as the peak rate $R_{\max}$ goes to infinity, which one of course would expect with infinite bandwidth available.

5.2.6 Outage-based capacity analysis

The worst-case design method used to compute the capacity η gives a rather coarse picture of the interference situation in a cellular communication system. In the analysis so far, a very simple propagation model was used. Also, in computing the capacity we have assumed the worst-case situation where the mobile terminal is at the border of the cell. Provided it is possible to design such an interference-free cell plan, all assignment failures will be caused by the fact that the number of channels in a cell is not sufficient.

In real cellular systems, the received signal levels are not smoothly distance dependent. On the contrary, due to shadow fading the received signal may fluctuate considerably. The straightforward approach to handling signal variation is to include a fade margin on top of the minimum required SINR. To guarantee that practically *all* terminals in this way obtain an adequate SINR would require impractically large reuse distances and thus uninterestingly low capacities. Any real cellular system will thus trade off a small fraction of terminals not satisfying the SINR requirement for a higher capacity.

For this purpose, we start by introducing the stochastic variable Q, denoting the number of terminals that have been assigned a channel but that cannot obtain an adequate SINR. From a user's perspective, having received a channel does no good if the channel turns out to be useless. Fast-moving terminals may have some hope that they move into a more favorable situation quickly as opposed to slowly moving terminals. In fact, in the latter case, it may not even be possible to distinguish between a conventional assignment failure and a poor channel. Communication is not possible in either of these cases. The performance measure can now be generalized, and the rate at which the system fails to assign *useful* channels is termed the *(total) assignment failure rate*, defined as:

$$\begin{aligned}\nu = \frac{E\{Z\}}{E\{M_c\}} = \frac{E\{\max(0, M_c - \eta)\} + E\{Q\}}{\omega A_c} \\ = \frac{E\{\max(0, M_c - \eta)\}}{\omega A_c} + \chi = \nu_p + \chi, \end{aligned} \tag{5.22}$$

where, as in the previous section, M_c is the number of active users and Z is the number of failures within the cell. The *primitive assignment failure rate* ν_p is thus nothing more than the assignment failure rate for a system design for co-channel interference-free conditions. Further, the *interference rate* was introduced,

$$\chi = \frac{E\{Q\}}{\omega A_c}, \tag{5.23}$$

which at moderate traffic loads ω will closely approximate the *outage probability*

$$\chi \approx \Pr(\Gamma < \gamma_t). \tag{5.24}$$

Considering the uplink of a hexagonal cellular system, the received SINR at the base station 0 at a distance r from the mobile terminal is written as

$$\Gamma_{\mathrm{UL}} = \frac{m_0 \frac{cP_0}{r^\alpha}}{\sum_{k=1}^{M} X_k m_k \frac{cP_k}{D_k^\alpha} + N} \approx \frac{1}{\sum_{k=1}^{M} X_k \frac{P_k}{P_0} \frac{m_k}{m_0} \left(\frac{r}{D_k}\right)^\alpha}, \tag{5.25}$$

where m_k is a log-normally distributed random variable with σ standard deviation modeling the shadow fading, N is the noise power, M is the total number of co-channel cells, P_k is the transmitter power of mobile terminal k, and X_k is the activity probability defined as

$$X_k = \begin{cases} 1, & \text{mobile in cell } k \text{ is active} \\ 0, & \text{otherwise.} \end{cases}$$

The quantity in (5.25) is a sum of a random number of independent log-normally distributed random variables. As the channel assignment in the different cells can be assumed to be independent, the number of terms in this sum will be binomially distributed with probability

$$\begin{aligned} q = \Pr(X_k = 1) &= \frac{E[\min(M_c, \eta)]}{\eta} \\ &= \sum_{k=0}^{\eta} \frac{k}{\eta} \frac{(\omega A_c)^k}{k!} e^{-\omega A_c} + \sum_{k=\eta+1}^{+\infty} \frac{(\omega A_c)^k}{k!} e^{-\omega A_c} \\ &= 1 - \sum_{k=0}^{\eta} \left(1 - \frac{k}{\eta}\right) \frac{(\omega A_c)^k}{k!} e^{-\omega A_c} = \varpi_\eta \left(1 - \nu_p\right), \end{aligned} \tag{5.26}$$

which is proportional to the relative traffic load and the primitive assignment failure rate. The parameter q is denoted as the *activity factor*.

A complication in the analysis is the fact that in most systems, the base station selection will be such that the instantaneously strongest base station is selected. Due to shadow fading, this is not necessarily the geographically closest base station to the mobile terminal. A far-away terminal is likely to create more damaging interference (or will be more susceptible to interference) than close-by terminals.

Exact calculation of the outage probability is not possible and one has to resort to numerical integrations, approximations or computer simulations. A possible approximation is to assume that the interference is dominated by the strongest interferer.

We further consider the six closest interferers, which are all at a distance of about $D = D_0\sqrt{3K}$ from the receiver. For a given number of active interferers l and a given distance r, the outage probability can then be approximated as

$$\Pr(\Gamma < \gamma_t | r, l) = \Pr\left(\sum_{k=1}^{l} \frac{P_k}{P_0}\frac{m_k}{m_0}\left(\frac{r}{D_k}\right)^{\alpha} > \frac{1}{\gamma_t}\right)$$

$$\approx 1 - \left[Q\left(\frac{10\alpha\log_{10}(r/D_0) - 10\log_{10}((3K)^{\alpha/2}/\gamma_t)}{\sigma\sqrt{2}}\right)\right]^{l},$$

where we have assumed the same transmitter power for all active terminals. Averaging over the distance r, the average probability for a given number of interferers we get is

$$\Pr(\Gamma < \gamma_t | l)$$
$$= 1 - \int_0^1 \left[Q\left(\frac{10\alpha\log_{10}(x) - 10\log_{10}((3K)^{\alpha/2}/\gamma_t)}{\sigma\sqrt{2}}\right)\right]^{l} 2x\,dx.$$

With an activity factor q, the receiver will experience interference from l interferers with probability

$$v_l = \binom{6}{l} q^l (1-q)^{6-l},$$

and the average outage probability in this case can be written as

$$\Pr(\Gamma < \gamma_t)$$
$$= 1 - \sum_{l=0}^{6} v_l \int_0^1 \left[Q\left(\frac{10\alpha\log_{10}(x) - 10\log_{10}((3K)^{\alpha/2}/\gamma_t)}{\sigma\sqrt{2}}\right)\right]^{l} 2x\,dx.$$

When modeling the cellular system as a birth–death process with finite waiting room, the total blocking probability (or Grade of Service, GoS) including the effect of fading channels can be written as

$$\mathrm{GoS} = E_\eta + \chi, \tag{5.27}$$

where E_η is the Erlang-B formula defined earlier.

Example 5.6: Consider the downlink of a wireless network with a total of $C = 100$ channels and a required SINR threshold for good signal quality of $\gamma_t = 20$ dB.

The path loss is log-normally distributed with a standard deviation of $\sigma = 6$ dB with a log-average proportional to $40\log_{10}(r)$, where r is the distance between the transmitter and the receiver. The additive noise can be neglected.

- Determine the number of channel groups at high traffic loads if we neglect the effect of fading (deterministic path loss).
- Determine the outage probability at high traffic loads for the K obtained in the deterministic case.

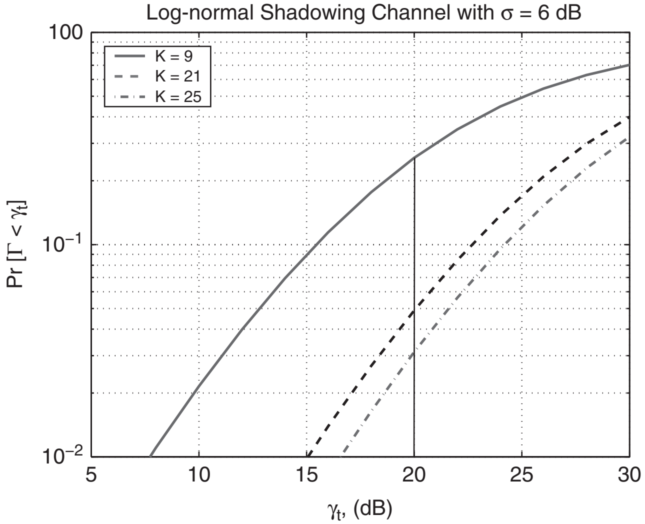


Figure 5.13 Outage probability of a cellular system for different cluster sizes in shadow fading channels.

- Determine the number of channel groups required to achieve an outage probability of at most 5% for the high load case.

Solution: Using the expression in (5.8) the cluster size is obtained as

$$K \geq \frac{1}{3}(6\gamma_t)^{2/\alpha} = \frac{1}{3}(600)^{1/2} = 8.2.$$

Checking the condition in (5.5) we see that the cluster size should be $K = 9$. This gives a capacity of $\eta = \frac{100}{9} = 11$ channels/cell.

Figure 5.13 illustrates the outage probability for different values of K where we see that for $K = 9$ and $\gamma_t = 20$ dB, the outage probability is $\Pr(\Gamma < \gamma_t) = 25\%$.

For an outage probability of 5%, we notice from the figure that the cluster size should be $K = 21$. This gives a capacity of $\eta = \frac{100}{21} = 4$ channels/cell, a reduction by a factor of 3 in comparison with the non-fading case.

Example 5.7: Design a mobile telephony system using $C = 400$ channel pairs. The modulation and coding requires an SIR of 13 dB to provide sufficient link quality. The path loss is log-normally distributed, standard deviation $\sigma = 6$ dB, with a log-average proportional to $40\log_{10}(r)$ where r is the distance between the transmitter and the receiver. The grade of service is defined by the combination of outage probability and assignment failure so that the total assignment failure rate is given by

$$\nu = \nu_p + \chi.$$

Determine the capacity of the system in calls/cell if the total assignment failure rate should not exceed 2% ($\nu \leq 2\%$).

Table 5.1 Outage probability for different values of K and $\gamma_t = 13$ dB.

K	χ
16	2%
19	1.3%
21	0.47%
25	0.25%

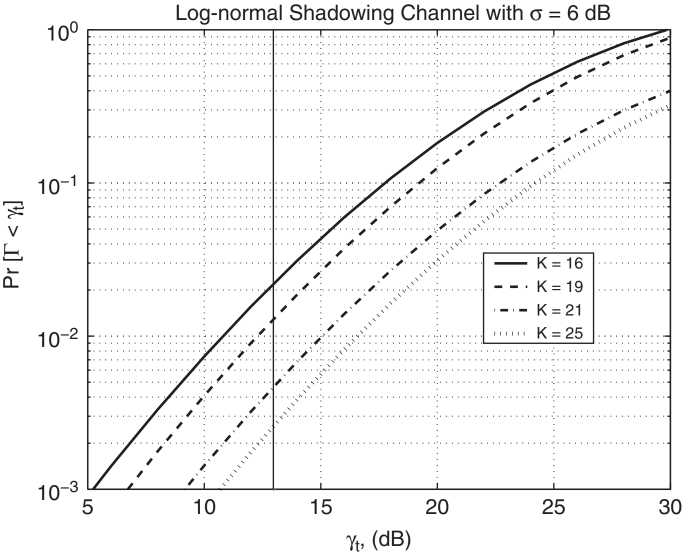


Figure 5.14 Outage probability of the system of Example 5.7.

Solution: We start by evaluating the activity factor q. First we assume $q = 1$ and compute χ for some different values of K as illustrated in Figure 5.14. The obtained outage probability results are presented in Table 5.1.

Locking the total assignment failure rate to the 2% level will yield the primitive assignment failure rate as

$$\nu_p = 2 - \chi \ (\%).$$

Step 1: Full load ($q = 1$): With full load, the assignment failure results of Table 5.2 are obtained. This approximate analysis indicates that there is a shallow maximum around $K = 19$ with a load of about 15 calls/cell. The activity factor q corresponding to this design would be

$$q = \varpi(1 - \nu_p) = \frac{15}{21}(1 - 0.7\%) = 0.7,$$

which is different from the value of q assumed in the calculations given in Table 5.2. Hence, we derive new outage probability graphs with this new activity probability.

Table 5.2 Assignment failure rate for the full load case (Step 1)

K	η	χ	ν_p	ν	ωA_c
16	25	2.0%	0.0%	2%	0
19	21	1.30%	0.70%	2%	15
21	29	0.47%	1.53%	2%	14
25	16	0.25%	1.75%	2%	12

Table 5.3 Assignment failure rate for the case $q = 0.7$ (Step 2)

K	η	χ	ν_p	ν	ωA_c
13	30	2.0%	0.0%	2%	0
16	25	0.75%	1.25%	2%	17
19	21	0.44%	1.56%	2%	15
21	19	0.32%	1.68%	2%	14

Step 2: True load ($q = 0.7$)**:** With this new load, the assignment failure rate results of Table 5.3 are obtained. From this table the cluster size $K = 16$ is found to be the best choice yielding a capacity of about 17 calls/cell.

With this cluster size, we compute the new activity factor as

$$q = \varpi \left(1 - \nu_p\right) = \frac{17}{25}\,(1 - 1.25\%) \approx= 0.7,$$

which is the same as the previous activity probability. With this convergence of a fixed activity probability we stop the iterations and the final value for the cluster size is $K = 16$ giving a capacity of 17 calls/cell.

In the analysis, the assumption has been made that the log-normal shadow fading variables G_i in (5.25) are all uncorrelated. This is a reasonable assumption for the uplink where the interfering signals reach the base station from mobiles in different positions (due to their uniform locations). For the downlink direction the situation is, however, somewhat different. The mobile will receive signals from a few very distinct directions. In addition, if a mobile terminal moves into some poor location (indoors or behind some large obstacle) many or all base stations will be shadowed more or less simultaneously. On the other hand, if the terminal moves to some high location with good reception, it is likely that many or even all base stations will be received with high received signal levels. To account for this correlation between the signals received from the different base stations, one may modify the shadow fading variables of (5.25) by introducing a very simple first order linear correlation model such as

$$Y_k = \sqrt{\rho}\; Y_M + \sqrt{1-\rho}\; Y_{kB}$$

$$G_k = 10^{Y_k/10}$$

where $Y_M, Y_{1B}, Y_{2B}, \ldots$ are independent Gaussian random variables. We note that for $\rho = 1$ all shadow fading variables are the same for all base stations.

5.2.7 Directional antennas and sectorizations

The systems that were analyzed in the previous section all used omni-directional antennas. This is a quite natural solution since one strives to achieve a good coverage of the service area. Further, a radio access point (RAP) usually has no information about where the mobile terminal is located. The disadvantage of an omni-directional antenna system is (besides providing low antenna gain) that the access port radiates interference power in all directions, not just in the wanted direction. If directional transmitting antennas could be employed, the interference levels could be substantially reduced. In addition, a directional receiving antenna would be able to suppress interference from unwanted directions. This could, in turn, result in lower reuse distances being used (i.e. higher capacity).

Directional antennas tend to be rather bulky and are for this reason most commonly used at the RAPs. The antennas can be *fixed* or *adaptive*. RAPs using fixed directional antennas will usually require an array of antennas, each covering a *sector* with the apex at the RAP. A mobile connected to the RAP moving in the service area is served by one of the directional antennas. As the mobile moves, a *handoff* to a different antenna may be necessary. An adaptive antenna array, on the other hand, would continuously track the mobile terminal. This has been shown to be a very effective tool both for mobile and fixed wireless systems. In particular, modern signal processing applied to electronically steerable array antennas has, in recent years, proven useful in this context. The latter lies beyond the scope of this book. The analysis here will be confined to the simpler, fixed antenna case. Most of the benefits will, however, appear in this case.

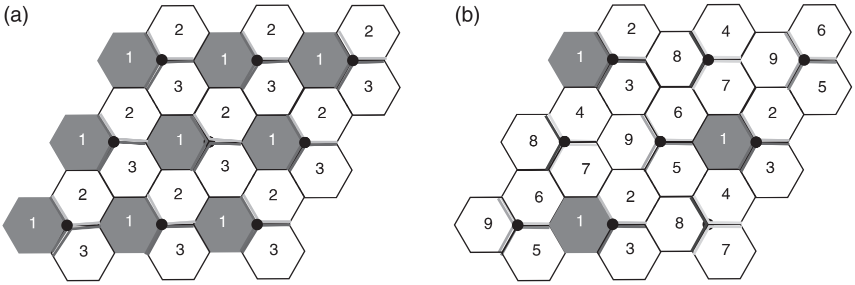
Assume that the base stations are equipped with ideal *sector antennas* which in the horizontal plane of the far-field have the relative intensity

$$S(\phi) \approx \begin{cases} \frac{2\pi}{\phi_h}, & |\phi| \leq \phi_h/2 \\ \frac{1}{A_{\text{sl}}}, & |\phi| > \phi_h/2 \end{cases} \tag{5.28}$$

The antenna radiation diagram [L. Ahlin et al., 2006] is characterized by the horizontal lobe width ϕ_h and the side lobe attenuation A_{sl}. Note that an omni-directional antenna would be described by $\phi_h = 2\pi$. Again, it is assumed that the RAP is located at the center of the cell and points (one of) its antenna in the direction $\phi = 0$. The uplink received SINR from a certain mobile terminal within cell 0 can now be written as

$$\Gamma_{\text{UL}} = \frac{\frac{c_t P_{k,0} S(\phi_0)}{r^\alpha}}{\sum_{i=1}^{M} X_i \frac{c_t P_{k,i} S(\phi_k)}{D_{k,i}^\alpha + N}}. \tag{5.29}$$

We assume that the terminal communicating with the base station is within the main lobe of the antenna. Signals from interfering mobile terminals (using the same channel) that are located outside the main lobe are (strongly) attenuated. However, signals from mobile terminals within the main lobe are received with higher power than in the



1/3 cluster pattern.

3/9 cluster pattern.

Figure 5.15 120 degree site patterns for 1/3 and 3/9 reuse in a hexagonal cellular system.

omni-directional case. However, since the desired signal experiences the same antenna gain, the relative interference power of these latter terminals remains the same as in the omni-directional case.

If the side lobe attenuation is large, and neglecting the noise, the expression in (5.25) can be rewritten as

$$\Gamma_{\mathrm{UL}} = \frac{m_0 \frac{cP_0}{r^\alpha}}{\sum_{k:|\phi_k|<\phi_h/2} X_k m_k \frac{cP_k}{D_k^\alpha} + N}. \tag{5.30}$$

Compared to the omni-directional case, the SINR increases in the same proportion as the effective number of interfering terminals (those falling into the main lobe). As a crude approximation, one can say that an antenna with main lobe width $2\pi/N$ reduces the average interference power by a factor N.

In practice, one is also interested in using directional antennas to reduce the number of base station sites and thus reduce the infrastructure costs of the system. The idea is to colocate three base stations at the same site, as illustrated in Figure 5.15. The site is located at the corner of three cells, where each base station uses a 120°-sector antenna to cover its cell. Otherwise, the cellular pattern remains the same. Thanks to the directional antennas the interference power is reduced since typically only $1/3$ ($= 120/360$) of the interferers are visible to the base station. It can be shown that for reuse factors divisible by three, it is possible to find frequency assignments where the nearest co-channel neighboring base stations are not facing each other (thus avoiding the worst source of interference). Such reuse patterns are denoted $x/3x$ where x denotes the reuse factor for number of base stations (i.e. how many different RAP channel assignments there are, whereas the $3x$ denotes the reuse factor for cells). Common examples are $1/3, 3/9$ and $4/12$. In the $1/3$ case, all base stations use all frequencies but split the channels into three sets, one for each of the sectors. On the other hand, placing a base station at the corner of a cell will cause a terminal to be up to twice as far from its base station as in the cell-center case. The following example analyzes such a system using simple analytical tools.

Example 5.8: Assume that the RAPs in the system in Example 5.2 use ideal 120° sector antennas, located at the corner of the cells. Further, assume that all base stations using a certain channel group illuminate their cells from the same direction. Estimate the capacity η that can be used if the other requirements are the same as in Example 5.2 with no fading, $C = 100$ channels, propagation path loss exponent $\alpha = 4$, and $\gamma_t = 20$ dB.

Solution: The worst interference case occurs when a terminal is at its largest distance from the RAP, i.e. at distance $2D_0$. The 120° sector antennas will receive interference from only two of the six nearest co-channel neighbors. Since all RAP antennas are turned the same way, both interfering terminals have to be located behind the serving RAP. We approximate the distance from these mobiles to the RAP under investigation as D. Considering only the first tier of interferers, the SIR can now be expressed as

$$\mathrm{SIR} \approx \frac{\frac{cP}{(2D_0)^4}}{2\frac{cP}{(D)^4}} = \frac{1}{2}\left(\frac{\Delta}{2}\right)^4 > 100,$$

which gives $\Delta > 2 \times 200^{1/4} \approx 7.52$.

With $\Delta = \sqrt{3K}$ we get

$$K = \frac{1}{3}\Delta^2 \approx 19, \quad \text{i.e. } K = 19.$$

The number of channels per cell is

$$\eta = \frac{M}{K} = \frac{100}{19} \approx 5 \text{ channels/cell},$$

which is slightly lower than that obtained with the omni-directional antennas case in Example 5.2. The number of RAP sites is, however, only 1/3 of the number used in the omni-directional case.

5.3 CDMA cellular systems

We have seen in the previous section that orthogonal channels are quite sensitive to co-channel interference and a cellular network will operate properly only with a high reuse distance. A high reuse distance ensures that the interference experienced by active users is limited and allows all active users to reach their quality of service in terms of bit error probability or data rate. However, a high reuse distance reduces the number of available channels in each cell, which limits the capacity of the system.

Let us reconsider the received signal-to-noise plus interference ratio for user i as given in the previous section,

$$\Gamma_i = \frac{G_{ii}P_i}{\sum_{k\neq i} G_{ki}P_k + N} \geq \gamma_t, \tag{5.31}$$

where P is the transmitted power, G_{ki} is the link gain between transmitter k and receiver i, N is the noise power, and γ_t is the required SINR threshold for good signal quality.

For digital communication signals, we often use the average energy per bit, compared to the total noise power spectral density, as a quality measure. With that, we obtain the received signal energy-to-total noise power spectral density for user i as

$$\left(\frac{E_b}{I_0}\right)_i = \frac{W}{R}\frac{G_{ii}P_i}{\sum_{k\neq i} G_{ki}P_k + N_0 W} = \frac{W}{R}\Gamma_i, \tag{5.32}$$

where $R = 1/T_s$ is the symbol data rate which is approximately equal to the bandwidth of user i, W is the interference bandwidth, and N_0 is the additive noise power spectral density.

Denoting the target $(E_b/I_0)_i$ that achieves the required quality of service by ξ_t we have

$$\frac{W}{R}\frac{G_{ii}P_i}{\sum_{k\neq i} G_{ki}P_k + N_0 W} \geq \xi_t \rightsquigarrow \Gamma_i \geq \xi_t \frac{R}{W} = \gamma_t. \tag{5.33}$$

Hence, we now have a relation between the two thresholds ξ_t and γ_t:

$$\gamma_t = \xi_t - 10\log_{10}\left(\frac{W}{R}\right), \quad \text{in dB.} \tag{5.34}$$

The quantity $10\log_{10}\left(\frac{W}{R}\right) - \xi_t$ is called the interference margin provided by the radio link. A positive interference margin means that it is possible to achieve the required quality of service even when the received signal power is lower than that of the interference.

For the orthogonal channels described in the previous section, we have $W \approx R$ and the interference margin is negative, which means that the received signal power has to be larger than that of the total interference to ensure the required quality of service. This explains the sensitivity of orthogonal channels to co-channel interference and the need for a large reuse distance to keep co-channel interference at a minimum. Hence, by increasing the bandwidth of the interfering users we can lower their power spectral density levels and make the radio links with the cellular system more robust to interference.

As an illustrative example consider the uplink capacity of orthogonal cellular systems and ignore the additive white noise. For a given threshold γ_t, the cluster size is obtained as

$$K \geq \left\lceil \frac{1}{3}\left(1 + [6\gamma_t]^{1/\alpha}\right)^2 \right\rceil = \left\lceil \frac{1}{3}\left(1 + \left[6\xi_t \frac{R}{W}\right]^{1/\alpha}\right)^2 \right\rceil. \tag{5.35}$$

It can be observed from the expression in (5.35) that the cluster size decreases with increasing W/R and will eventually reach 1 (see Figure 5.16). Hence, with a proper selection of the interference margin, it is possible to design wireless networks with a reuse distance equal to *one* (universal reuse distance) where all the channels are used in all cells. It is clear that with universal reuse ($K = 1$), there is no need for frequency planning since the same bandwidth will be used in all cells. Direct sequence code division multiple access (DS-CDMA) cellular systems employ pseudorandom (PN) spreading codes to widen the bandwidth of the interference and reject most of the co-channel interference. The ratio W/R is referred to as the processing gain in DS-CDMA systems.

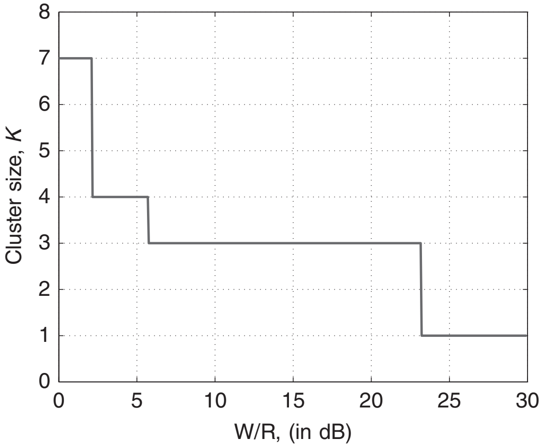


Figure 5.16 Cluster size as a function of the interference margin.

DS-CDMA is based on the principle of using a high interference margin to reduce the effects of external interference and improve the capacity of the system, and was originally used in military applications for secure communications. In the 1980s CDMA was considered as an access scheme in cellular systems and the US IS-95 cellular CDMA standard [Standard-95 TIA/EIA Interim, 1993] was created. DS-CDMA is characterized by signals with a bandwidth much larger than the information bandwidth $1/T_s$. DS modulation employs pseudorandom waveforms to spread the information signal over the total spreading bandwidth. Commercial DS-CDMA cellular systems employ two-layered spreading codes. This multiple spreading code allocation provides flexible system deployment and operation. In fact, multiple spreading codes make it possible to provide waveform orthogonality among all users of the same cell while maintaining mutual randomness between users of different cells. Orthogonality can be achieved through the channelization code layer, a set of orthogonal short spreading codes such as the variable-length Walsh orthogonal sequence set where each cell uses the same set of orthogonal codes. A long scrambling code is employed as a second layer to reduce external interference (inter-cell interference). A cell-specific scrambling code (common to all users) is employed in the downlink and a user-specific code on the uplink. Hence, each transmission is characterized by the combination of a channelization code and a scrambling code. Figure 5.17 illustrates the coding structure used in IS-95 and WCDMA for the downlink case [E. H. Dinan and B. Jabbari, 1998]. The IS-95 standard employs inter-cell synchronous scrambling codes where each cell uses a different time offset of the same pseudorandom code while the WCDMA standard employs inter-cell asynchronous scrambling codes where each cell uses a different pseudorandom code. Table 5.4 gives the type of spreading codes used in the IS-95 and WCDMA cellular systems. On the uplink, all terminals employ the same very long sequence with each terminal identified by a specific time offset in the IS-95 standard while each terminal employs a different scrambling code in the WCDMA standard.

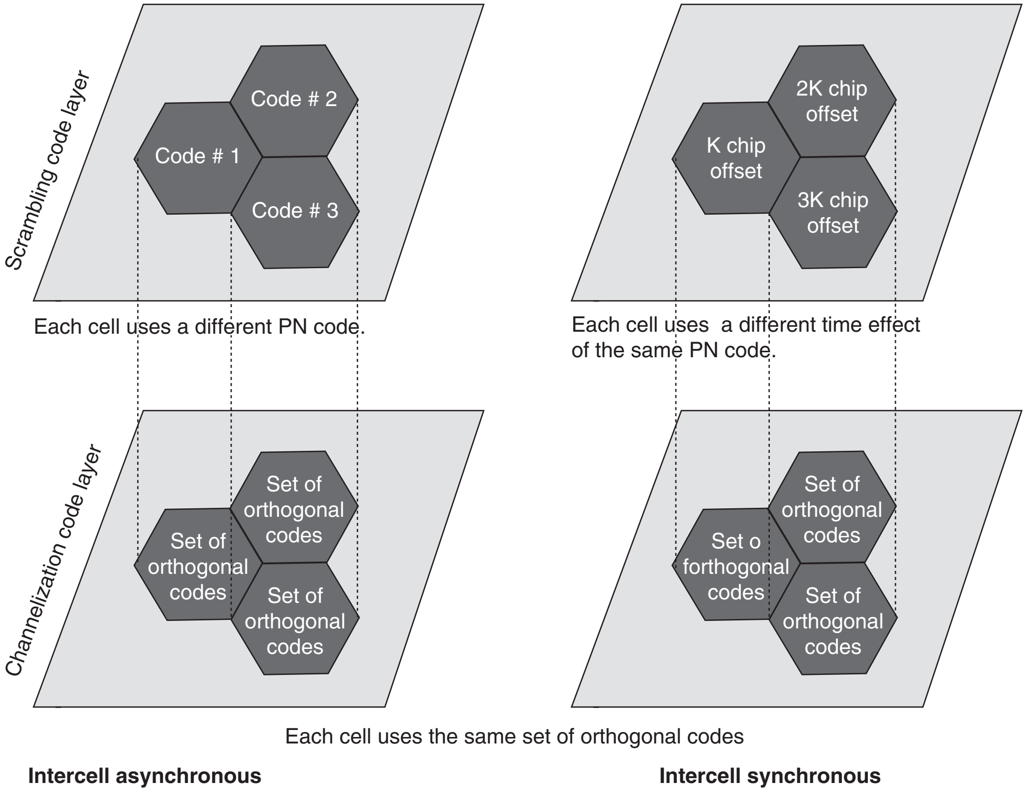


Figure 5.17 Inter-cell synchronous and asynchronous spreading code assignment [E. H. Dinan and B. Jabbari, 1998].

Even though DS-CDMA employs orthogonal channelization codes, it is quite difficult to keep signals within one cell perfectly orthogonal due to the presence of fading multipath channels and non-ideal synchronization between users within the same cell. This is particularly true for the uplink case where users are at different positions within the cell and their transmitted signals will multiplex in the air before getting to the base station. Hence, DS-CDMA systems will experience both intra-cell interference and inter-cell interference.

5.3.1 Uplink capacity of DS-CDMA systems

Consider a DS-CDMA system with M active users in the cell of interest. Without loss of generality, assume the randomly picked cell is cell 1. The received signal at the base station receiver of cell 1 can be written as follows:

$$r(t) = \sum_{i=1}^{M} y_{1,i}(t) + \sum_{k=2}^{L} z_k(t) + n(t), \tag{5.36}$$

where $y_{1,i}(t)$ is the received signal from mobile terminal i within cell 1, $z_k(t)$ is the received signal from all active mobile terminals within cell k, and $n(t)$ is the additive

Table 5.4 Spreading codes used in IS-95 and WCDMA cellular systems

	CDMA (IS-95)		WCDMA	
	Forward link	Reverse link	Forward link	Reverse link
Channelization code	Walsh orthogonal sequences of length 64.	-	Variable-length orthogonal sequences.	Variable-length orthogonal sequences.
Scrambling code	Different offsets of an m-sequence with period $2^{15}-1$. A common PN for all users of a cell.	Different offsets of an m-sequence with period $2^{15}-1$. A common PN for all users of a cell.	Gold sequences with period $2^{18}-1$. A common PN for users of a cell.	Very large set of Kasami sequence. Gold sequences with period $2^{41}-1$ optional.

white Gaussian noise. Assuming single user detection and a quality of service defined by a required signal energy-to-interference-plus-power-spectral-density threshold ξ_t, the signal energy-to-interference-plus-noise-power-spectral-density ratio for user m can be written as

$$\left(\frac{E_b}{I_0}\right)_m = \frac{W}{R}\frac{P_{1,m}}{\sum_{i=1,i\neq m}^{M} P_{1,i} + \sum_{k=2}^{L} P'_k + N_0 W} \tag{5.37}$$

$$\geq \xi_t, \quad 1 \leq m \leq M \tag{5.38}$$

where $P_{1,i}$ is the received power from mobile terminal i within the cell, L is the total number of base stations within the service area, P'_k is the received power from all active mobile terminals within cell k, R is the user data rate, and W is the spreading bandwidth.

Single-cell capacity

For an isolated single-cell DS-CDMA system, the received signal-energy-to-interference-plus-noise-power-spectral-density ratio becomes

$$\left(\frac{E_b}{I_0}\right)_m = \frac{W}{R}\frac{P_{1,m}}{\sum_{i=1,i\neq m}^{M} P_{1,i} + N_0 W} \tag{5.39}$$

$$\geq \xi_t, \quad m = 1,2,\ldots,M \tag{5.40}$$

Since all users within the same cell interfere with each other, it is important to control their transmit power so that they will not jam each other. A good power control procedure is to ensure that the received power from the user at the base station receiver is constant regardless of its position within the cell. Such a procedure allows users to communicate without generating excessive interference. Assuming ideal constant received power control (CRPC), the received E_b/I_0 becomes

$$\left(\frac{E_b}{I_0}\right)_m = \frac{W}{R}\frac{P_1}{(M-1)P_1+N_0W} \tag{5.41}$$

$$\geq \xi_t, \quad m=1,2,\ldots,M \tag{5.42}$$

where P_1 is the constant received power from a mobile terminal at the base station receiver of cell 1.

The inequality in the expression of (5.42) will be satisfied if and only if

$$M \leq 1+\frac{W}{R}\left(\frac{1}{\xi_t}-\frac{1}{\xi_0}\right), \tag{5.43}$$

where $\xi_0 = \frac{P_1}{N_0R}$ is the received signal-to-noise ratio at the base station receiver.

It can be seen from (5.43) that the number of supported mobile terminals in the cell is proportional to the processing gain W/R and is inversely proportional to the required threshold ξ_t. It is also important to notice that the received SNR ξ_0 has to be larger than the required threshold ξ_t to ensure a number of users larger than one.

Example 5.9: The IS-95 system has a user data rate of 9.6 kbps and a spreading bandwidth $W = 1.25$ MHz. The required signal energy to interference plus noise power spectral density that ensures a bit error rate of 10^{-3} (standard voice communication systems) is $\xi_t = 7$ dB [K. S. Gilhousen et al., 1991]. With these parameters we may compute an upper bound on the maximum number of mobile terminals that can be supported from (5.43) to get

$$M \leq 1+\frac{1.25\text{ MHz}}{9.6\text{ kbps}}\left(\frac{1}{5.01}-\frac{1}{\xi_0}\right) \leq 26\text{ users/cell}. \tag{5.44}$$

The capacity expression given in (5.43) has been obtained under an ideal power control scheme. However, in practice, the power control is not ideal and the corresponding received E_b/I_0 becomes random. The capacity M can then be determined only statistically, i.e. M is computed for a given outage probability requirement.

Multi-cell capacity

With multiple cells, each mobile-terminal-to-base-station-receiver link will experience interference from mobile terminals within its cell and interference from mobile terminals in neighboring cells. With constant received power control, the received signal energy-to-interference-plus-noise-power-spectral-density ratio at the base station receiver of cell 1 for mobile terminal m is given by

$$\begin{aligned}\left(\frac{E_b}{I_0}\right)_m &= \frac{W}{R}\frac{P_1}{(M-1)P_1+\sum_{k=2}^{L}P'_k+N_0W}\\ &= \frac{W}{R}\frac{P_1}{I_{\text{intra}}+I_{\text{inter}}+N_0W}.\end{aligned} \tag{5.45}$$

Denoting by L the total number of cells and the cell of interest by cell 1, the received E_b/I_0 can be rewritten as

$$\left(\frac{E_b}{I_0}\right)_m = \frac{W}{R}\frac{P_1}{(M-1)P_1 + \sum_{l=2}^{L}\sum_{k=1}^{M}\frac{G_{l,1}}{G_{l,k}}P_l + N_0 W}, \tag{5.46}$$

where $G_{k,l}$ is the link gain between mobile terminal l and base station k, P_l is the constant received power at base station l, and M is the number of active mobile terminals per cell.

DS-CDMA systems employ soft handover where each active user is always connected to its best base station, i.e. the radio link toward the base station that gives the best link gain. In this case, we have

$$\frac{G_{l,1}}{G_{l,k}} \leq 1 \quad \forall\, k, \tag{5.47}$$

because otherwise the mobile terminal would switch to the cell site for which the attenuation is minimized.

The capacity of multi-cell CDMA is then obtained as

$$\eta = \{M \,|P\left[(E_b/I_0)_k < \xi_t\right] = p_o\}, \tag{5.48}$$

where p_o is the required outage probability of the system. The outage probability can be computed numerically, as in [K. S. Gilhousen et al., 1991], or by means of computer simulations.

A rough estimate of multi-cell CDMA capacity can be obtained by considering the average of the received $(E_b/I_0)_k$. Hence, one can write

$$E\left\{(E_b/I_0)_k\right\} \approx \frac{W}{R}\frac{P_1}{(M-1)P_1 + E\{I_{\text{inter}}\} + N_0 W}. \tag{5.49}$$

The total interference power received by base station 1 from the active mobile terminals within the neighboring cells can be approximated as

$$E\{I_{\text{inter}}\} \approx f E\{I_{\text{intra}}\} = f(M-1)P_1. \tag{5.50}$$

Using the approximation of (5.50) in (5.49) and solving for M we get

$$M = 1 + \frac{1}{F}\frac{W}{R}\left(\frac{1}{\xi_t} - \frac{1}{\xi_0}\right), \tag{5.51}$$

where we have defined F as [A. J. Viterbi et al., 1994a]

$$F = 1 + f = \frac{\text{Total interference power}}{\text{Own cell interference power}}. \tag{5.52}$$

This is known as the F-factor and is defined as the ratio of the total interference power at the base station to the home cell interference power.

The F-factor has been investigated extensively in the literature. The results obtained show that most of the interference in CDMA systems comes from within the cell and the external interference is just a portion of that interference. The value of F depends on the propagation conditions of the radio channel. For CDMA systems without shadowing, the other-cell interference factor, f, was found to be 0.44 for a propagation exponent of

$\alpha = 4$ [A. J. Viterbi et al., 1994b]. For CDMA systems with shadowing, f was found to be 0.55 for a propagation exponent of $\alpha = 4$ and a log-normal shadowing standard deviation of $\sigma = 8$ dB [A. J. Viterbi et al., 1994a]. In general, f is a function of α and σ. It also depends on the soft handover parameters such as the active base station set.

Several methods can be used to increase the capacity of CDMA systems, for instance reducing ξ_t through improved channel coding or increasing the total bandwidth of the system (increasing the processing gain). As seen from (5.51), the capacity of CDMA is inversely proportional to the factor F. Hence, any traffic variations in the neighboring cells will affect the number of users that can be served within the cell. The capacity of CDMA systems is then not limited by the available number of spreading codes but rather by the amount of interference experienced. Interference management is key in CDMA cellular systems.

Sectorization and voice activity monitoring

The most fundamental aspect associated with tuning CDMA networks is managing interference levels. As the capacity of CDMA systems is a function of interference, any variation in interference across the network translates directly into a variation in the system capacity. Spatial isolation through directional antennas is one way of reducing interference in CDMA systems. Consider an example where three directional antennas having 120° effective beam widths are employed. Now the interference sources seen by any of these antennas are approximately one-third of those seen by the omni-directional antenna. This reduces the interference term in the denominator of (5.49) by a factor of approximately 3 and consequently the number of users (M) in (5.51) is increased by approximately the same factor. Denoting this sectorization factor by n_{sec}, the CDMA capacity becomes

$$M = 1 + \frac{n_{\text{sec}}}{F}\frac{W}{R}\left(\frac{1}{\xi_t} - \frac{1}{\xi_0}\right). \tag{5.53}$$

Voice activity monitoring is a feature present in most digital vocoder where the transmission is suppressed for that user when no voice is present. Modeling the talk spurt and silence as exponential with mean τ_1 and τ_2 seconds, respectively, the voice activity factor can be written as follows:

$$q = \frac{\tau_1}{\tau_1 + \tau_2}. \tag{5.54}$$

The transmission rate of a given mobile i can be modeled as a stochastic process with an average transmission rate given by

$$R = qR_{ts} + (1-q)R_{ss}, \tag{5.55}$$

where R_{ts} is the rate of the talk spurt and R_{ss} is the rate of the silence spurt. Hence, the corresponding average received power at the base station from such a user with voice activity detection becomes qP_i, where P_i is the average received power without voice activity detection.

With voice activity monitoring, the experienced average interference at the base station is reduced by a factor $1/q$ and the multiple-cell capacity becomes

$$M = 1 + \frac{n_{\mathrm{sec}}}{Fq}\frac{W}{R}\left(\frac{1}{\xi_t} - \frac{1}{\xi_0}\right). \tag{5.56}$$

The above expression assumes that all the interference comes from a single interferer with voice activity factor q. In fact, since users are independent, the interference due to the other users is a binomially distributed random variable. This variation in interference should be taken into account to obtain a more accurate estimation of the CDMA capacity. The capacity would then be evaluated for a given outage probability.

Dimensioning the uplink power budget in CDMA is based on the link quality equation. From (5.56), the required transmit power of mobile user m can be written as

$$P_{t,k} = \frac{P_1}{G_{1,k}} \geq \frac{N_0 W}{G_{1,k}\left(\frac{Fq}{n_{\mathrm{sec}}} + \frac{W}{R}\frac{1}{\xi_t}\right)} \frac{1}{1 - \frac{M}{1+\frac{n_{\mathrm{sec}}}{Fq}\frac{W}{R}\frac{1}{\xi_t}}}, \tag{5.57}$$

where $P_{t,i}$ is the transmit power of user i and $G_{1,i}$ is the link gain between user i and the base station receiver of cell one. Here the transmit power is determined by the noise power $N_0 W$, the required threshold ξ_t, the link gain $G_{1,i}$, and the processing gain W/R.

From (5.57), we define the uplink loading as

$$L = \frac{M}{1 + \frac{n_{\mathrm{sec}}}{Fq}\frac{W}{R}\frac{1}{\xi_t}}. \tag{5.58}$$

When L approaches 1, the uplink capacity approaches its maximum, *the pole capacity*, where the required transmit power of the mobile user approaches infinity. Hence, an upper bound on the uplink capacity of CDMA can be written as

$$M \leq M_{\mathrm{pole}} = 1 + \frac{n_{\mathrm{sec}}}{Fq}\frac{W}{R}\frac{1}{\xi_t}, \tag{5.59}$$

where M_{pole} denotes the pole capacity.

Real systems operate below the pole capacity. This means that, in order to provide service to users, there should be enough power available to maintain an acceptable quality of service. For the uplink, this means that the total received interference power at the base station receiver must not be too high. From (5.46), the total received uplink power at the base station receiver can be written as

$$I_{\mathrm{tot}} = MP_1 + \sum_{l=2}^{L}\sum_{k=1}^{M}\frac{G_{l,1}}{G_{l,k}}P_l + N_0 W, \tag{5.60}$$

with M representing the total number of active users per cell. We define the uplink noise rise as the total received power divided by the background noise power with

$$\eta = \frac{I_{\mathrm{tot}}}{N_0 W} = \frac{1}{1-L}, \tag{5.61}$$

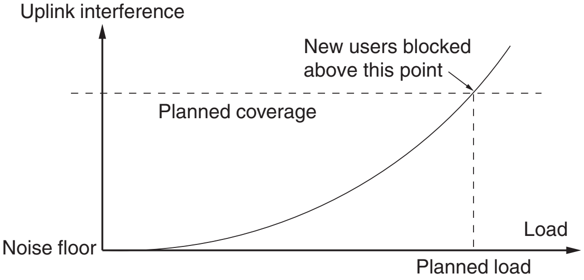


Figure 5.18 Total received interference as a function of the relative load L.

which clearly shows that $L = 0$ implies $I_{\text{tot}} = N_0 W$, i.e., an empty CDMA system with background noise only. As L approaches the system is operated close to the pole capacity and the interference power goes to infinity (see Figure 5.18).

To account for the effects of imperfect power control and the dynamics of the system, a CDMA system should not run too close to the pole capacity. Running too close to the pole capacity will make the system unstable and mobile users at the border of the cell will not be able to communicate. In fact, in CDMA systems, one can relate the coverage to the number of active users where coverage can be traded off for capacity and vice versa. From the expression in (5.57), the transmit power of user i is obtained as

$$P_{t,i} = \min\left\{P_{t,\max}, \frac{n_{\text{sec}} N_0 W}{G_{1,k} F q} \frac{1}{M_{\text{pole}} - M}\right\}, \tag{5.62}$$

where $P_{t,\max}$ is the maximum transmit power of the mobile user.

It is clear from (5.62) that, by allowing the loading to increase, then users at the border of the cell can no longer reach the required target ξ_t and, as a result, the cell shrinks due to loading. Conversely, if mobile users are transmitting at their maximum power, loading should not be permitted to increase in order not to affect the coverage of the cell. This phenomenon couples coverage and loading in CDMA systems where one can trade off coverage for capacity and vice versa. In general, loading of 50% to 75% is an appropriate compromise between loading and coverage.

5.3.2 Traffic-based capacity of DS-CDMA systems

In general, the number of channels (spreading codes) in CDMA systems is very large and much larger than the pole capacity obtained in (5.59). Hence, blocking in CDMA systems occurs due to bad links and not due to lack of available codes. Assuming that the number of available codes is very large, each cell in CDMA systems can be modeled as a M/G/∞ queue. In cellular networks, calls can be modeled as a Poisson process with λ calls/s and the call duration can be assumed to be exponentially distributed with average duration $1/\mu$ seconds. Hence, for a single cell with no external interference, the number of active users is a random variable with

$$\Pr(M = n) = \frac{(\lambda/\mu)^n}{n!} e^{-\lambda/\mu} = \frac{\rho^n}{n!} e^{-\rho}, \tag{5.63}$$

where $\rho = \lambda/\mu$ is the traffic load.

Since the load in CDMA systems is directly related to the experienced interference, the effect of voice activity detection and space isolation with sector antennas can be modeled by modifying the traffic of the cell as

$$\rho = \frac{q}{n_{\text{sec}}} \frac{\lambda}{\mu}. \tag{5.64}$$

Furthermore, since the inter-cell interference has been modeled as a fraction of intra-cell interference, the multiple cell case can also be seen as a single cell with a higher traffic load due to the higher interference experienced. Hence, the number of active mobile users in multiple cell CDMA systems is modeled as a Poisson random variable with probability density function

$$\Pr(M = n) = \frac{[(1+f)\lambda/\mu]^n}{n!} e^{-(1+f)\frac{\lambda}{\mu}}. \tag{5.65}$$

With both voice activity detection and sector antennas, the cell load is weighted with the factor q/n_{sec} in the above expression.

Having M active users within the cell and assuming perfect constant received power control, the received signal-energy-to-interference-plus-noise-power-spectral-density ratio for user m can be written as

$$\left(\frac{E_b}{I_0}\right)_m = \frac{W}{R} \frac{P_1}{(1+f)(M-1)P_1 + N_0 W} \geq \xi_t. \tag{5.66}$$

For a given planned load L, an assignment failure occurs when the total number of active users within the cell exceeds LM_{pole} as shown in Figure 5.18. Hence, the blocking probability can be written as

$$\nu = \Pr(M \geq K_0), \quad \text{with } K_0 = \lfloor LM_{\text{pole}} \rfloor, \tag{5.67}$$

where $\lfloor x \rfloor$ denotes the greatest integer function of a real number x.

Solving for ν we get

$$\nu = e^{-\rho} \sum_{n=K_0}^{+\infty} \frac{\rho^n}{n!}, \tag{5.68}$$

where

$$\rho = \begin{cases} \frac{\lambda}{\mu}, & \text{single cell w/ omni antenna} \\ (1+f)\frac{\lambda}{\mu}, & \text{multi-cell (m-cell) w/ omni antenna} \\ (1+f)\frac{q}{n_{\text{sec}}}\frac{\lambda}{\mu}, & \text{m-cell w/ voice and sector antenna} \end{cases} \tag{5.69}$$

and

$$M_{\text{pole}} = \begin{cases} 1 + \frac{W}{R}\frac{1}{\xi_t}, & \text{single cell w/ omni antenna} \\ 1 + \frac{1}{1+f}\frac{W}{R}\frac{1}{\xi_t}, & \text{m-cell w/ omni antenna} \\ 1 + \frac{n_{\text{sec}}}{(1+f)q}\frac{W}{R}\frac{1}{\xi_t}, & \text{m-cell w/ voice and sector antenna} \end{cases} \tag{5.70}$$

5.3.3 Downlink capacity of DS-CDMA systems

For the downlink, the capacity depends on the downlink transmit power, which is limited by the base station power amplifier. The process takes the form of a power allocation where each user uses part of the base station transmit power. The objective is of course to ensure the required quality of service for users within the cell. Hence, users near the border of the cell may need more power than users close to the base station. Denoting by ϕ_i the fraction of power allocated to user i within the cell, then the total number of mobile users in the downlink should satisfy the following:

$$\sum_{i=1}^{M} \phi_i \leq 1. \tag{5.71}$$

Due to intra- and inter-cell interference, the more users we have in the cell, the stronger the power required for each individual user. This requires that the mobile users measure their received signal to interference plus noise ratios. Also, each user is connected to their best base station. For instance, if we have a total of B base stations within the service area, the mobile user connects to the base station from which it receives the strongest power. Denoting by 1 the best base station for user m, the received signal-energy-to-noise-power-spectral-density ratio can be written as

$$\begin{aligned}\left(\frac{E_b}{I_0}\right)_m &= \frac{W}{R}\frac{\phi_m P_{1,m}}{\theta_m(1-\phi_m)P_{1,m}+\sum_{k=2}^{B} P_{k,m}+N_0 W} \\ &= \frac{W}{R}\frac{\phi_m P G_{1,m}}{\theta_m(1-\phi_m)PG_{1,m}+\sum_{k=2}^{B} PG_{k,m}+N_0 W} \qquad (5.72) \\ &\geq \xi_t \qquad (5.73)\end{aligned}$$

where θ_m (with $0 \leq \theta_m \leq 1$) denotes the orthogonality factor between the user waveforms within the same cell, which depends on multipath propagation conditions ($\theta_m = 0$ corresponds to full orthogonality in the case of a single propagation path), $P_{k,m}$ is the received power from base station k. P is the base station transmit power (assumed the same for all base stations) and $G_{i,m}$ is the link gain between base station i and the mobile user m. As mobile user m is connected to the strongest base station, assumed base station 1 in (5.73), we have $P_{1,m} > P_{k,m}$, $\forall\ k \geq 2$.

The fraction of power needed for mobile m can be obtained from (5.73) and is written as

$$\phi_m \geq \frac{1}{\theta_m + \frac{W}{R}\frac{1}{\xi_t}}\left(\theta_m + f_{\mathrm{DL},m} + \frac{N_0 W}{P_{1,m}}\right), \tag{5.74}$$

where we have defined the interference factor $f_{\mathrm{DL},m}$ as

$$f_{\mathrm{DL},m} = \sum_{k=2}^{M} \frac{P_{k,m}}{P_{1,m}} \tag{5.75}$$

which represents the other-to-own cell received power ratio for connection m at the position of the corresponding mobile user.

Combining the inequality in (5.71) with the expression in (5.74) we get

$$\sum_{m=1}^{M}\left(\theta_m + f_{\mathrm{DL},m} + \frac{N_0 W}{PG_{1,m}}\right) \leq \sum_{m=1}^{M}\phi_m\left(\theta_m + \frac{W}{R}\frac{1}{\xi_t}\right). \tag{5.76}$$

Defining the average other-to-own-cell interference ratio in the downlink as $f_{\mathrm{DL}} = \frac{1}{M}\sum_{i=1}^{M} f_{\mathrm{DL},m}$, the average orthogonality factor as $\theta = \frac{1}{M}\sum_{i=1}^{M}\theta_m$, and $\theta' = \sum_{i=1}^{M}\phi_m\theta_m \approx \theta$, the expression in (5.76) becomes

$$M\left(\theta + f_{\mathrm{DL}}\right) + \frac{N_0 W}{P}\sum_{m=1}^{M}\frac{1}{G_{1,m}} \leq \theta' + \frac{W}{R}\frac{1}{\xi_t}. \tag{5.77}$$

Solving for the total downlink base station transmission power we get

$$P \geq \frac{1}{\theta + \frac{W}{R}\frac{1}{\xi_t}}\frac{\sum_{m=1}^{M}\frac{N_0 W}{G_{1,m}}}{1 - M\frac{\theta + f_{\mathrm{DL}}}{\theta + \frac{W}{R}\frac{1}{\xi_t}}}. \tag{5.78}$$

It is convenient to define the downlink loading, L_{DL}, as

$$L_{\mathrm{DL}} = M\frac{\theta + f_{\mathrm{DL}}}{\theta + \frac{W}{R}\frac{1}{\xi_t}}, \tag{5.79}$$

and the total downlink base station transmission power becomes

$$P \geq \frac{1}{\theta + \frac{W}{R}\frac{1}{\xi_t}}\frac{\sum_{m=1}^{M}\frac{N_0 W}{G_{1,m}}}{1 - L_{\mathrm{DL}}}. \tag{5.80}$$

When L_{DL} approaches 1, the download capacity approaches its maximum, the pole capacity, where the required transmit power approaches infinity. From (5.80), we can write an upper bound on the downlink capacity of cellular CDMA systems as

$$M \leq \frac{\theta + \frac{W}{R}\frac{1}{\xi_t}}{\theta + f_{\mathrm{DL}}} \tag{5.81}$$

with θ being the orthogonality factor between users within the same cell.

5.3.4 Multi-service DS-CDMA systems

We have limited our discussion in the previous sections to single-service CDMA systems. However, direct sequence CDMA systems can support a wide variety of communication services, such as voice, video, multimedia, and circuit- and packet-mode data communication. In a DS-CDMA system, the radio resource can be allocated to the users by regulating their transmission power, modulation scheme, channel coding, and spreading gains (transmission rates) such that their QoS is satisfied. Each service is, in general, characterized by a different QoS.

Consider the uplink of a multi-service CDMA system and assume that we have a set $M = \{1, 2, \ldots, M\}$ active users within the cell, say cell one. Denote the set of available rates within the CDMA systems as $\{R_{t,1}, R_{t,2}, \ldots, R_{t,K}\}$ with the corresponding E_b/I_0 targets $\{\xi_{t,1},\xi_{t,2},\ldots,\xi_{t,K}\}$. At a given time slot, the received signal-energy-to-interference-plus-noise-power-spectral-density ratio for user i can be written as

$$\left(\frac{E_b}{I_0}\right)_i = \frac{W}{R_i}\frac{P_{t,i}G_{1,i}}{\sum_{k=1,k\neq i}^{M} P_{t,k}G_{1,i} + I_{\text{inter}} + N_0 W}, \tag{5.82}$$

$$i = 1,2,\ldots,M \tag{5.83}$$

where $R_i \in \{R_{t,1},R_{t,2},\ldots,R_{t,K}\}$ is the data rate of user i, $P_{t,i}$ is its transmission power, and $G_{1,i}$ is the link gain between mobile user i and the base station receiver.

We assume that the required quality of service is reflected through the received E_b/I_0. Hence, controlling the transmission powers and spreading gains of the users amounts to directly controlling the QoS measures and link efficiency. With ξ_i denoting the required target E_b/I_0 for mobile user i that corresponds to the data rate R_i, a set of linear equations is obtained from (5.83) and is written as

$$\left(\frac{E_b}{I_0}\right)_i = \frac{W}{R_i}\frac{P_{t,i}G_{1,i}}{\sum_{k=1,k\neq i}^{M} P_{t,k}G_{1,i} + I_{\text{inter}} + N_0 W} \geq \xi_i, \tag{5.84}$$

$$i = 1,2,\ldots,M \tag{5.85}$$

Solving the above set of linear equations, an expression for the required transmission power of mobile user i is written as

$$P_{t,i} = \frac{1}{G_{1,i}}\frac{1}{1+\frac{W}{R_i}\frac{1}{\xi_i}}\frac{I_{\text{inter}} + N_0 W}{1 - \sum_{k=1}^{M}\left(\frac{1}{1+\frac{W}{R_k}\frac{1}{\xi_k}}\right)}. \tag{5.86}$$

Similar to the single service case, we define the loading factor for multi-service CDMA as

$$L = \sum_{k=1}^{M}\left(\frac{1}{1+\frac{W}{R_k}\frac{1}{\xi_k}}\right). \tag{5.87}$$

As L approaches 1, the required transmit power approaches infinity. Hence, the required quality of service is satisfied for M active users if and only if the following inequality is satisfied:

$$\sum_{k=1}^{M}\left(\frac{1}{1+\frac{W}{R_k}\frac{1}{\xi_k}}\right) < 1. \tag{5.88}$$

However, in practice, the transmission power of each mobile user is limited and hence the expression in (5.86) is bounded as

$$0 \leq P_{t,i} = \frac{1}{G_{1,i}}\frac{1}{1+\frac{W}{R_i}\frac{1}{\xi_i}}\frac{I_{\text{inter}} + N_0 W}{1 - \sum_{k=1}^{M}\left(\frac{1}{1+\frac{W}{R_k}\frac{1}{\xi_k}}\right)} \leq P_{t,\max}. \tag{5.89}$$

Solving for the loading factor we get

$$\sum_{k=1}^{M}\left(\frac{1}{1+\frac{W}{R_k}\frac{1}{\xi_k}}\right) \leq 1-\frac{I_{\text{inter}}+N_0W}{\min\limits_{1\leq i\leq M}\left[P_{t,\max}G_{1,i}\left(1+\dfrac{W}{R_i}\dfrac{1}{\xi_i}\right)\right]}. \tag{5.90}$$

It can be seen that the loading in (5.90) is limited by the connection having the lowest link gain. Hence, in a DS-CDMA with best-effort services, the total cell throughput is maximized by allocating the higher data rates to the users having the best link gains. With such rate allocation, the term $\min\limits_{1\leq i\leq M}\left[G_{1,i}\left(1+\dfrac{W}{R_i}\dfrac{1}{\xi_i}\right)\right]$ is maximized and is directly translated into a higher cell throughput.

For the downlink case, the received signal-energy-to-interference-power-spectral-density ratio in a multi-service CDMA can be written as

$$\left(\frac{E_b}{I_0}\right)_m = \frac{W}{R_m}\frac{\phi_m PG_{1,m}}{\theta_m(1-\phi_m)PG_{1,m}+\sum_{k=2}^{B}PG_{k,m}+N_0W} \geq \xi_i \tag{5.91}$$

with $\sum_{m=1}^{M}\phi_m \leq 1$, $\xi_m \in \{\xi_{t,1},\xi_{t,2},\ldots,\xi_{t,K}\}$, and $R_m \in \{R_{t,1},R_{t,2},\ldots,R_{t,K}\}$.

Solving for the fraction of power needed for mobile user m that ensures its quality of service we get

$$\phi_m \geq \frac{1}{\theta_m+\frac{W}{R_m}\frac{1}{\xi_m}}\left(\theta_m+f_{\text{DL},m}+\frac{N_0W}{PG_{1,m}}\right), \tag{5.92}$$

which is inversely proportional to the link gain $G_{1,m}$, and proportional to the QoS indicator $R_m\xi_m$ and the interference factor $f_{\text{DL},m}$. Hence, in a multi-service DS-CDMA with best-effort services, higher rates should be allocated to mobile users having the higher link gains (closer to the base station).

Exercises

5.1 Consider a cellular system with a total of 120 channels. The system is modeled by a regular hexagonal grid with the base stations placed in the corners of each cell. The base stations use directional antennas. Assume that the base stations use ideal sectorized antennas with 60° lobe width, i.e. every cell is divided into two halves requiring two channel groups. The path loss exponent is 4. The SIR requirement for good signal quality is 15 dB and thermal noise can be neglected. Compare the capacity (measured as channels/unit area) of this system with the two following systems:

a) A system with 120° sector antennas.
b) A system with omni-directional antennas and base stations in the center of the cell.

5.2 One way to improve the capacity of cellular systems is to employ a two-channel bandwidth scheme, where a hexagonal cell is divided into two concentric hexagons as shown in Figure 5.19. The inner hexagon is serviced by 15 kHz channels, while the outer is serviced by 30 kHz channels. Suppose that the 30 kHz channels require an

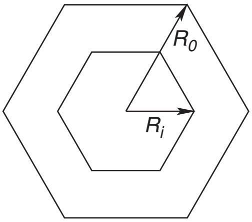


Figure 5.19 Geometry of Problem 5.2.

18 dB threshold to maintain acceptable radio link quality, while the 15 kHz channels require 24 dB. The path loss exponent is 4.

a) Consider the downlink and assume a fourth-law path loss model. Determine the ratios, $\Delta_0 = D/R_0$ and $\Delta_i = D/R_i$.
b) Determine the ratio of the inner and outer hexagonal areas, A_i/A_0.
c) Let N_i and N_0 be the number of users allocated to the inner and outer portions of the cell respectively, and assume that the channels are allocated such that $N_i/N_0 = A_i/(A_0 - A_i)$. Determine the increase in capacity (channel per cell) over a conventional one-channel bandwidth system that uses only 30 kHz channels, if the system uses a 7-cell reuse cluster.

Consider only the first tier of interference and neglect the thermal noise.

5.3 A cellular telephony system uses a static channel allocation with $\eta = 25$ channels per cell. The arrival process of new calls is modeled as a Poisson process with arrival rate $\lambda = 180$ calls/h/km^2. Every call has an exponentially distributed duration with an average $1/\mu = 2.5$ minutes.

a) Determine the traffic intensity (erlang/cell) if the cell radius is $R = 1$ km.
b) What is the blocking probability at this traffic load?
c) Determine the channel assignment failure rate at the traffic load if we can assume that (almost) all calls are handled by the system.

5.4 A wireless cellular system has a total of 100 channels with all base stations placed on a highway. The base stations are equidistant with a distance of 2 km between two neighboring base stations. The base stations employ directional antennas and radiate power in one direction only. Assume that the propagation loss increases with the fourth power of the distance and the thermal noise can be neglected.

a) Determine the capacity (in channels/km) that can be achieved if the SIR requirement is 20 dB.
b) Determine the blocking probability as a function of the relative traffic load (erlang/cell/channel).
c) Determine the channel assignment failure rate as a function of the relative load (active terminals/cell/channel).

5.5 Consider a hexagonal cellular system where base stations have been mounted in the middle of the cells. Each base station uses an omni-directional antenna and an equivalent isotropically radiated transmit power P_t. We have a total of 90 channels available and are aiming at 30 channels/cell. The required SIR for good signal quality is $\gamma_t = 8$ dB and the propagation path loss exponent is $\alpha = 4$. The transmitted power at the base station has been adjusted such that the SNR at the border of the cell is 10 dB.

a) Considering the downlink, compute the outage probability of this cellular system.
b) Determine the required extra power margin needed that ensures full coverage (zero outage probability).

5.6 A wireless cellular system is modeled by a regular hexagonal grid with base stations placed in the middle of the cells with omni-directional antennas. The distance between two neighboring base stations is 2 km. The total number of channels owned by the operator is 400 channels. The modulation scheme requires a minimum SIR of 15 dB to ensure the required quality of service. Assuming that the traffic is uniformly distributed, what is the capacity of the system (in erlang/km^2)

a) if a blocking probability of 5% and an outage probability of 5% can be accepted?
b) if the GoS, defined as the sum of outage and blocking probabilities, is not to exceed 10%?

The operator using a design according to b) is now forced to surrender half of its bandwidth to competing digital services, i.e. only 200 channels remain.

What is the capacity in the system now?

5.7 A one-dimensional cellular system has 80 channels at its disposal. The modulation scheme employed requires a minimum SIR of 14 dB. Determine the minimum distance between radio access ports if the channel assignment failure rate is not to exceed 1% and the required area capacity is 5 mobiles/km. Assume that the path loss increases as the $\alpha = 2.5$ power of the distance.

Hint: Make and motivate necessary approximations.

5.8 A wireless cellular system is planned to cover a straight highway with base stations placed along it. Each base station has 40 channels at its disposal and employs omni-directional antennas. A modulation and coding scheme requiring 19 dB SIR is employed. The required area capacity is 10 mobiles/km for a channel assignment failure rate of 2%. Assume that the propagation loss increases with the fourth power of the distance.

a) Assume a shadow fading channel with a log-normal standard deviation $\sigma = 6$ dB. The fading correlation is angle-dependent and is at its maximum when the angle between the links is zero. Determine the outage probability.
b) How many fewer base stations are needed if we use directional antennas (radiating only in one direction along the road) instead of omni-directional antennas (radiating in all directions)?

5.9 In a mobile data system a $K = 4$ reuse pattern is used. The transmit power in the base station is adjusted such that the interference and noise components are equally large.

a) What is the SNR at the cell boundary?

b) What is the average cell capacity if a total of 20 MHz are used?

5.10 To offer broadband data access a mobile operator is planning to deploy a wireless network. Market surveys show that customers expect

- a bit rate of 400 kbps with 95% availability in the downlink (DL), and
- a bit rate of 200 kbps with 85% availability in the uplink (UL).

Base stations have a transmit power of 10 W and an antenna gain of 3 dB, while mobile terminals are associated with an equivalent isotropically radiated power of 20 dBm. Measurements show that the SNR at a distance r [km] from the base station can be written as

$$\gamma(r) = \frac{5}{r^{3.5}}, \quad \text{(i.e. for DL transmission).}$$

In both UL and DL, orthogonal channels are used and each channel has a bandwidth of $W = 1000$ kHz.

The achievable link bit rate R is given by the Shannon bound adjusted for efficiency losses, i.e.

$$R = W \log_2\left(1 + \frac{\Gamma}{3}\right) \quad \text{kbps,}$$

where Γ is the SINR at the receiver.

Assume that a hexagonal cell plan with a cell radius $r_{\text{cell}} = 500$ m is used.

a) Compute the cluster size, K_D, for the DL transmission that ensures the rate and coverage requirements.

b) Compute the cluster size, K_U, needed in the UL transmission.

c) Consider the DL transmission and assume that the SINR threshold is 3 dB higher than the one derived in part a) and that the cluster size remains the same. What is the base station density [BS/km^2] necessary to meet these new requirements?

5.11 Consider the uplink of a DS-CDMA system with base stations having omni-directional antennas and employing perfect received power control and voice activity detection with a voice activity factor $q = 0.5$. The number of users within the system is Poisson distributed with a mean of 8 users/cell. Inter-cell interference can be considered to be about 60% of own cell interference. The system is designed to operate properly at a load $L = 50\%$.

a) Determine the pole capacity M_p of this system if the assignment failure rate should not exceed 20%.

b) Determine the assignment failure rate if the CDMA system employs directional antennas with the antenna diagram shown in Figure 5.20 and base stations placed on the corners of the cells.

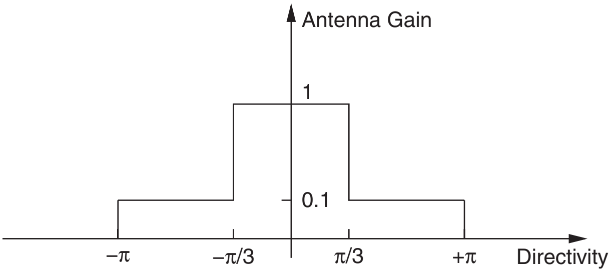


Figure 5.20 Antenna diagram.

5.12 Consider the uplink of a CDMA system with a service with processing gain PG = 20 dB and target $E_b/I_0 = 9.95$ dB. The number of users in the cell is Poisson distributed with mean $\rho = 20$. For a given instant, the probability that a randomly chosen user is active is $\nu = 0.5$. Perfect power control (constant received power control) can be assumed. Consider the following outage definition: a user is in outage if the interference amount is 10 times greater than the background noise amount.

a) Determine the outage probability for a single-cell system.
b) Consider a multiple-cell system where the interference from other cells is 50% of the own cell interference. Determine the outage probability in this case.
c) Assume that the user activity is reduced to $\nu' = 0.25$. Consider a single-cell system and determine the maximum E_b/I_0 for which the same outage probability is achieved as in a).

5.13 Consider a single-cell DS-CDMA system. The number of mobile terminals within the cell has a Poisson distribution with average ωA_c. The required SIR target $\gamma = -11.46$ dB should be reached with 95% probability.

a) What is the largest average number of terminals in the cell?
b) What happens in the cases when the SIR requirement cannot be met?
c) What is the pole capacity of this system, M_{pole}? Can you deduce this expression directly from the SIR expression?
d) Given that the (instantaneous) number of terminals in the cell is Poisson distributed with average $\rho = \omega A_c$ and that the pole capacity is M_{pole}, give an expression for the probability that the number of terminals exceeds M_{pole}.

5.14 The Erlang capacity evaluated for DS-CDMA systems was obtained assuming perfect constant received power control and a single rate for all users. An assignment failure was declared when the total interference I_0 exceeded the background noise level N_0 by an amount $1/\eta$. Now consider a multi-rate DS-CDMA system where each mobile terminal has its individual data rate, denoted R_i, and power control that provides each mobile terminal with its required $(E_b/I_0) = \xi_i$, when feasible.

a) Determine an expression for the blocking probability in this case.
b) Compare the obtained results with the single-rate case when $R_i = R$, $\forall\, i$.

5.15 As illustrated in Figure 5.21, consider a hexagonal cellular system with omnidirectional base stations and a static channel allocation scheme. The average SNR at the cell corner is 40 dB. In order to deliver a certain service the SINR has to be better than $\gamma = 13$ dB.

a) The operator can use 128 channels and has decided to use the following path loss model:

$$L_{\text{dB}} = 48 + 40\log_{10}(r)[m]$$

If we are aiming for at least 5 users/km^2, estimate how many base stations are needed to cover a region of 1000 km^2. Assume all transmitters have the same power, use a downlink calculation and reasonable simplifications.

b) Now choose a stochastic model for the SINR,

$$\Gamma = G\gamma_0,$$

where γ_0 is the average SINR and G is log-normal with expectation value 0 dB and standard deviation 4 dB. We want 98% availability (downlink) at the cell corner. How will this affect the number of base stations needed?

The information in Figure 5.22 can be used for the calculations.

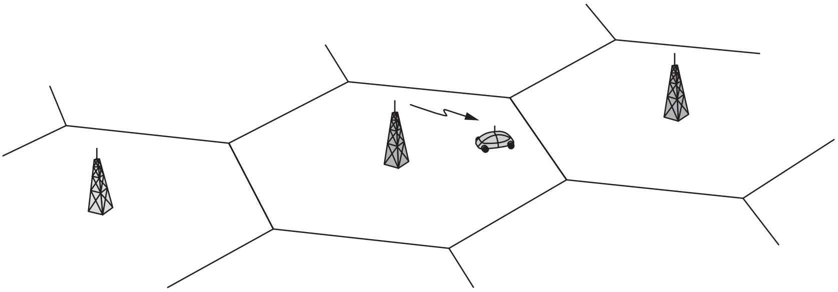
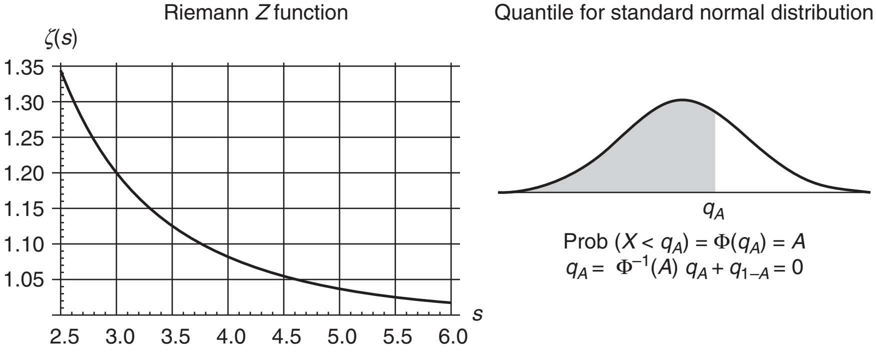


Figure 5.21 Cell illustration.



A	0.9	0.95	0.975	0.98	0.99	0.995	0.999	0.9995	0.9999	0.99995	0.99999
q_A	1.2816	1.6449	1.9600	2.0537	2.3263	2.5758	3.0902	3.2905	3.7190	3.8906	4.2649

Figure 5.22 Calculation table.

5.16 Consider the uplink of a single-cell CDMA system with a base station having an omni-directional antenna placed in the middle of the cell. The mobile users are uniformly distributed over the cell area. The required SINR for good signal quality is $\gamma_t = -19.96$ dB. The propagation path loss is modeled as

$$Lp(r) = 76 + 40\log_{10}(r) \quad \text{dB},$$

where r, in km, is the distance to the base station.

The terminals are using an ideal constant received power control, the maximum transmission power is $p_{\max} = 23$ dBm, and the noise power is $N_0 W = -33$ dBm.

a) Determine the outage probability for a user placed at a distance r from the base station as a function of the number of active users M and r.
b) Derive a relation showing the tradeoff between capacity (number of users) and coverage r, and make a plot of the obtained results.

5.17 In order to provide radio communication at a horse-race event, a base station is placed in the middle of the arena, so that any spectator is at exactly $d = 500$ m from the base station. However, each spectator experiences a log-normal distributed fading with standard deviation $\sigma = 5.102$ dB; assume that the fading values experienced by any two spectators are uncorrelated. On the uplink, the system uses CDMA with ideal SIR-target power control and provides two types of services. The voice service requires an SIR target $\gamma_v = -20$ dB, while the data service requires $\gamma_d = -14$ dB. One out of five users uses data services. If only one terminal were active, transmitting with the maximum allowed power, the median would be $\gamma_0 = 6$ dB.

a) Find an expression for the largest $\frac{1}{\eta}$ so that the users have access to the data service with probability 97.5%.
b) How many users can be served in this single-cell system under the conditions at point a)?

5.18 Consider the uplink in a single hexagonal cell CDMA system, with a cell radius $R = 1075$ m. The users are uniformly distributed over the service area with 7 users/km^2. The required signal-energy-to-noise-power-spectral-density ratio for good signal quality is $(E_b/N_0) = 7$ dB and the system processing gain is 128.

The path gain is modeled as $G(d) = c_0/d^{\alpha}$, with d in meters, $c_0 = -30$ dB, $\alpha = 3.5$, and the noise power is $N = -103$ dBm.

a) If ideal constant-received-power control is used, determine the minimum required received power that ensures the desired (E_b/N_0) target for all the users.
b) If the maximum transmitted power is limited to $P_{\max} = 100$ mW, determine the number of users that cannot reach the required signal quality.
c) How many users can the system support such that good signal quality is ensured for all the users anywhere within the cell?

5.19 Consider the uplink in a single-cell CDMA system, with a cell radius $\rho = 1000$ m. There are two service types in the system, corresponding to two different end-user bit rates. Both services require the same demodulated signal quality $E_b/N_0 = 5$ dB, but use different processing gains $P_{g,1} = 21$ dB and $P_{g,2} = 25$ dB (roughly corresponding to a 38 kbps data service and a 12.2 kbps voice service respectively). Each terminal uses ideal (perfect) power control to reach the desired SIR target; the maximum transmission power is $P_{\max} = 100$ mW.

a) What is the highest possible system load allowed in the system in order to provide a high bit rate service to a user located at the cell border? (Express it as load L, η or P_{tot}/N). The path gain is modeled as $G(d) = c_0/d^{\alpha}$, with $c_0 = -24$ dB and $\alpha = 3.5$, and the noise power is $N = -103$ dBm.
b) Assume that the cell contains only one user with high bit rate service and a number of M users with low bit rate service. What is the largest number M of low bit rate users for which the high bit rate user can be served at the cell border?
c) Repeat the computations and find M when the processing gain for the high bit rate service is $P_{g,1} = 14$ dB.

5.20 A WCDMA system with a chip rate of 3.84 Mcps and an information data rate of 9.6 kbps serves two kinds of users: *gold users* who require a bit-energy-to-interference-spectral-density ratio threshold of 10 dB and *silver users* who require a bit-energy-to-interference-spectral-density ratio threshold of only 7 dB.

a) Consider the uplink situation in a single-cell system and denote by x the number of gold users and by y the number of silver users. Derive the required received power, at the base station, from each terminal as a function of x and y.
b) Show that the capacity of this WCDMA system, denoted by $\eta = x + y$, is a surface in the (x, y) signal space. Make a plot of this signal space and identify the capacity region for which the signal quality can be maintained for all active users.

5.21 Consider the uplink of a DS-CDMA multi-cellular system where we have a total of N users per cell, a total bandwidth of W hertz, and a user data rate of R bits/s. The base station receiver employs nonlinear successive cancelation where users are detected successively and canceled from the received signal one by one. The strongest user is detected first, remodulated, and then canceled from the received signal. The obtained signal is then used to detect the next strongest user and so on.

The required bit-energy-to-interference-spectral-density ratio for good signal quality is denoted ξ_t and is the same for all users.

a) Assuming that there is no error propagation along the stages, compute the required powers for the different users that achieve the required threshold for all the users as a function of ξ_t, W, R, N_0 and I_e (inter-cell interference).
b) Derive an upper bound on the system capacity, N, if the total received power from the users within the cell is not to exceed 3 dB above the noise level $N_0 W + I_e$.

5.22 A WCDMA system with a chip rate of 3.84 Mcps supports two types of users. A low rate type with a data rate of 9.6 kbps and a high rate type with a data rate of 480 kbps. To provide these rates, the system uses orthogonal variable length (OVL) spreading codes as channelization codes and long random codes as scrambling codes.

The required bit-energy-to-interference-spectral-density ratio for good signal quality is denoted ξ_t and is the same for both user types

a) Consider the uplink situation in a single-cell system and denote by x the number of low rate users and by y the number of high rate users. Derive the required received power, at the base station, from each terminal as a function of x, y and ξ_t such that the SIR threshold is satisfied for all users.
b) Show that the capacity of this WCDMA system, denoted by $\eta = x + y$, is a surface in the (x, y) signal space. Make a plot of this signal space and identify the capacity region for which the signal quality can be maintained for all active users when $\xi_t =$ 10 dB.

5.23 We would like to plan a CDMA cellular system (we are interested in the uplink case). The considered DS-CDMA system has a processing gain (spreading factor) of $\frac{W}{R_s} = 256$, an information data rate of $Rs = 10$ kbps, and employs a perfect constant received power control. The maximum transmitted power of the mobile unit is $P_{\max} = 24$ dBm and the noise power spectral density is $N_0 = -197$ dBW/Hz. The propagation path loss is given by $PL = 37 + 40\log_{10}(r)$ dB where r is the distance between the transmitter and the receiver in meters. The required signal-energy-to-noise-power-spectral-density ratio for good signal quality is $\xi_t = 3$ dB.

a) Assuming hexagonal cells, determine the area capacity if the number of users per cell is $M = 30$ users/cell.
b) Repeat part a) if the number of users per cell is $M = 60$ users/cell and compare.

References

L. Ahlin, J. Zander and S. Ben Slimane. 2006. *Principles of Wireless Communications*. Lund: Studentlitteratur.

D. C. Cox. 1982. Cochannel interference considerations in frequency reuse small coverage area radio systems. *IEEE Trans. Commun.*, 30(1), 135–142.

E. H. Dinan and B. Jabbari. 1998. Spreading codes for direct sequence CDMA and wideband cdma cellular networks. *IEEE Communication Magazine*, 36(9), 48–54.

K. S. Gilhousen, I. M. Jacobs, R. Padovani, A. J. Viterbi, L. A. Weaver and C. E. Wheatley III. 1991. On the capacity of a cellular CDMA system. *IEEE Transactions on Vehicular Technology*, 40(2), 303–312.

W. C. Y. Lee. 1993. *Mobile Communication Fundamentals*. Wiley: New York.

R. A. Leese. 1997. A unified approach to the assignment of radio channels on a regular hexagonal grid. *IEEE Trans. Veh. Tech.*, 46(4), 968–980.

V. H. MacDonald. 1979. The cellular concept. *Bell Syst. Tech. J.*, 58(1), 15–42.

Standard-95 TIA/EIA Interim. 1993. *Mobile station–base station compatibility standard for dual-mode wideband spread spectrum cellular systems*. Tech. rept. Telecommunications Industry Association.

A. J. Viterbi, A. M. Viterbi and E. Zehavi. 1994a. Other-cell interference in cellular power-controlled CDMA. *IEEE Trans. Commun.*, 42, 1501–1504.

A. J. Viterbi, A. M. Viterbi, K. S. Gilhousen and E. Zehavi. 1994b. Soft handoff extends CDMA cell coverage and increases reverse link capacity. *IEEE Journal on Selected Areas in Communications*, 12(8), 1281–1288.

6 Transmitter power control

6.1 Introduction

In Chapters 3 and 4, we have introduced MAC and scheduling in wireless networks. To simplify the discussion, we have neglected the transmitter power that a terminal or access port should use after being granted channel access. Transmitter power control refers to techniques that determine the amounts of transmission power in wireless networks. If only one terminal is transmitting, the terminal can simply use any amount of power available to achieve the performance it desires. However, in a wireless system, the transmitter power control is more complicated because of the broadcasting nature of wireless communications. The transmitter power affects both the link quality and the interference environment in the network. Adjusting the transmitter power values of different links to improve the network performance is not a trivial problem. While increasing the transmitter power of a terminal increases its SIR, it also increases the interference in the other links in the system, causing these terminals to increase their power. Reducing the transmitter power decreases the interference with the other links, but jeopardizes the quality of its own link. In this chapter, we study transmitter power control for wireless systems.

6.2 Performance metric and conditions of achievability

In this section, we consider the links that are communicating on an arbitrary orthogonal channel, c_0, that is static at the time it is being observed. An example is the uplink communications in cellular networks, where the transmitters are the scheduled mobile terminals and the receivers are the corresponding access ports at the base stations. The roles are reversed for the downlink communications. The link gain between receiver i and transmitter j is given by g_{ij}. Transmitter and receiver i and j are logical entities and may denote the same physical one. There are Q transmitters on the channel and transmitter j uses transmission power P_j. The transmission powers of all the transmitters are denoted by the following vector:

$$\mathbf{P} = (P_1, P_2, \ldots, P_Q)^{\mathrm{T}}.$$

The interference from all transmitters on channel c_0 should be considered. The signal to interference plus noise ratio (SINR) in the receiver i can be derived as:

$$\Gamma_i = \frac{g_{ii}P_i}{\sum_{j=1,j\neq i}^{Q} g_{ij}P_j + n_i}, \tag{6.1}$$

where n_i denotes thermal noise power at receiver i.

DEFINITION 6.1 *Transmitter i is supported if its SINR satisfies*

$$\Gamma_i \geq \gamma_0, \tag{6.2}$$

where γ_0 is a target threshold.

Replace (6.1) in (6.2) and we have

$$P_i \geq \gamma_0\Big(\sum_{j=1,j\neq i}^{Q} \frac{g_{ij}}{g_{ii}}P_j + \frac{n_i}{g_{ii}}\Big), \tag{6.3}$$

which shows the minimal transmission power needed by transmitter i to achieve the target SIR, given the powers of the other transmitters.

Example 6.1: Minimum transmission power Two terminals are sending data to their corresponding access ports. The link gain matrix **G** is given by:

$$\mathbf{G} = \begin{pmatrix} 0.33 & 0.05 \\ 0.06 & 0.38 \end{pmatrix}$$

The target SIR γ_0 is 6 dB and the receiver noise power is 0.1 W. What transmission powers should the two terminals use to minimize their power consumption?

Solution: According to (6.3), the supporting conditions are illustrated in Figure 6.1. When the transmission powers are in region A, both terminals will be supported. When

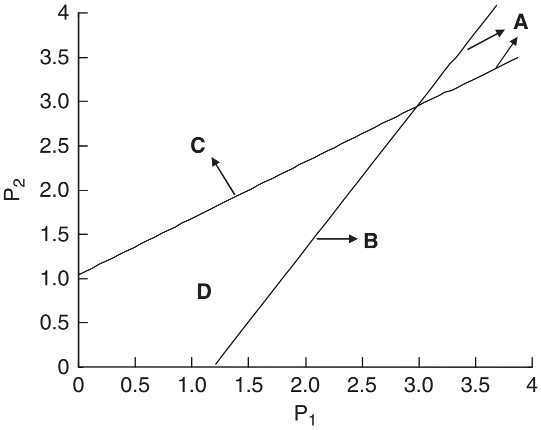


Figure 6.1 Two-link case.

the powers are in either B or C, either terminal 1 or 2 can be supported. No terminal will be supported in region D. The power vector in region A that minimizes the power consumption is the crossing point of the two lines and the one that satisfies the equation in (6.3). Solving the equations, we have the desired transmission powers $(P_1, P_2) = (2.96, 2.91)$.

One design objective is to maximize the number of supported terminals with properly selected transmission powers. In other words, we need to find a power vector that maximizes the number of supported users. Define the $Q \times Q$ *normalized* link gain matrix $\mathbf{H} = (h_{ij})$ such that

$$h_{ij} = \gamma_0 \frac{g_{ij}}{g_{ii}}$$

for $i \neq j$ and

$$h_{ij} = 0$$

for $i = j$, and the *normalized* noise vector $\eta = (\eta_i)$ such that

$$\eta_i = \gamma_0 \frac{n_i}{g_{ii}}.$$

We can rewrite the linear inequality in (6.3) as

$$P_i \geq \sum_{j=1}^{Q} (h_{ij} P_j + \eta_i).$$

The Q linear inequalities, $\Gamma_i \geq \gamma_0$, for all i, can now be described as

$$(\mathbf{I} - \mathbf{H})\mathbf{P} \geq \eta, \tag{6.4}$$

where $\mathbf{I}$ is the identity matrix. $\mathbf{A} \leq \mathbf{B}$ means that each component in $\mathbf{A}$ is no bigger than the corresponding one in $\mathbf{B}$. The power vector $\mathbf{P}$ is non-negative.

DEFINITION 6.2 *The target SINR γ_0 is achievable if there exists a non-negative power vector $\mathbf{P}$ such that for every i, $\Gamma_i \geq \gamma_0$.*

If there is a non-negative solution for (6.4), the SINR γ_0 is achievable. Using well-known linear algebra properties, we can derive the following proposition.

PROPOSITION 6.1 *There exists an achievable target SINR γ_0 when the dominant (largest) eigenvalue of matrix $\mathbf{H}$, denoted by $\rho(\mathbf{H})$, is no larger than one. When $\rho(\mathbf{H}) = 1$, γ_0 is achievable only with zero thermal noise.*

Proof There exists an achievable γ_0 only when there is a non-negative power vector $\mathbf{P}^*$ satisfying

$$(\mathbf{I} - \mathbf{H})\mathbf{P}^* = \eta. \tag{6.5}$$

When $\rho(\mathbf{H}) < 1$, the inverse matrix exists and

$$(\mathbf{I} - \mathbf{H})^{-1} = \sum_{k=0}^{\infty} \mathbf{H}^k,$$

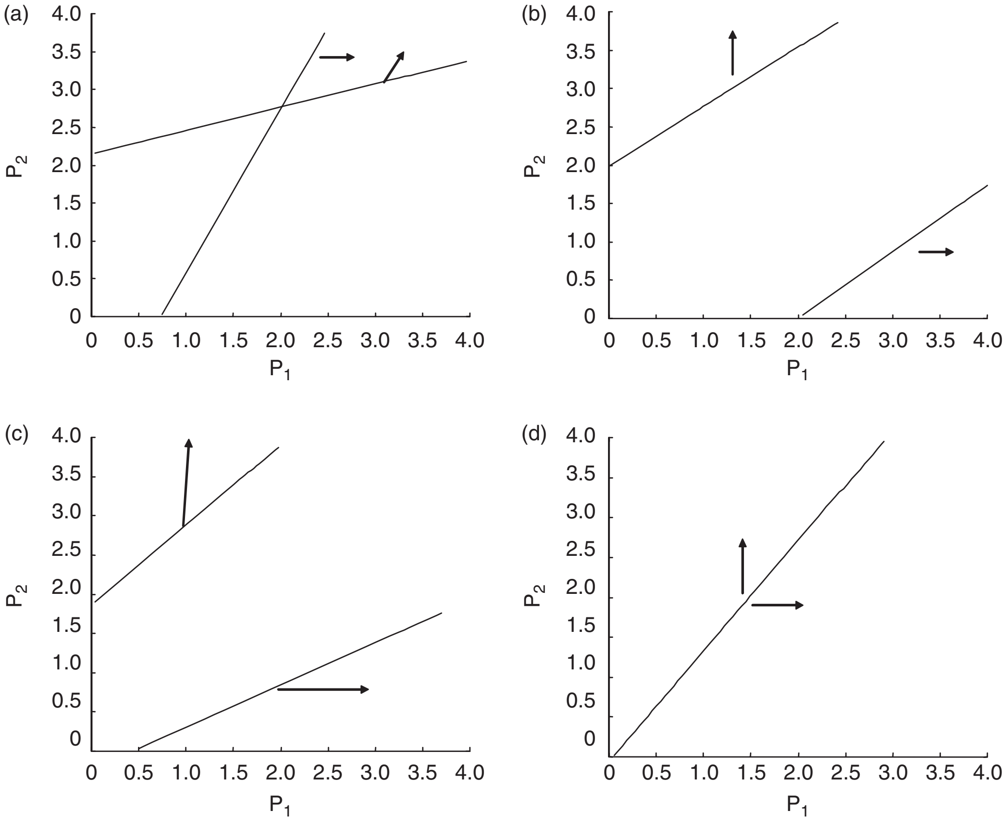


Figure 6.2 (a) $\rho(\mathbf{H}) < 1$ with $\eta > 0$; (b) $\rho(\mathbf{H}) = 1$ with $\eta > 0$; (c) $\rho(\mathbf{H}) > 1$ with $\eta > 0$; (d) $\rho(\mathbf{H}) = 1$ with $\eta = 0$.

so

$$(\mathbf{I} - \mathbf{H})^{-1} > 0$$

[R. S. Varga, 1962]. Then we have $\mathbf{P}^* = (\mathbf{I} - \mathbf{H})^{-1}\eta \geq 0$. In the case $\rho(\mathbf{H}) = 1$, $(\mathbf{I} - \mathbf{H})$ is singular and only when $\eta = 0$ will there be any non-negative solution to (6.5).

$\rho(\mathbf{H})$ increases with any entry of $\mathbf{H}$. When either the target SINR or link gain g_{ij} increases, it gets more difficult to achieve the target SINR γ_0. Proposition 6.1 is further illustrated in Figure 6.2 assuming two links. Both users can be supported in cases (a) and (d) while only one in cases (b) and (c). In the following sections, we focus on cases like (a) and (d). Cases like (b) and (c) will be discussed in Sections 6.3.2 and 6.4.5. □

6.3 Centralized power control

In this section, we discuss centralized power control assuming the channel information is known and the transmission powers can be determined by a central scheduler.

6.3.1 SIR balancing

In this section, assume the objective is to maximize the minimum SINR of all links. This can be achieved by equalizing the SINR of each transmitter, i.e. SINR balancing, and maximizing the balanced SINR. Assume the noiseless case, $\eta = 0$, to simplify the analysis. Based on (6.4), we have:

$$(\mathbf{I} - \mathbf{H})\mathbf{P} \geq 0.$$

To determine the maximum achievable γ_0, define matrix $\mathbf{A}$ such that $\mathbf{H} = \gamma_0 \mathbf{A}$. Let $\rho(\mathbf{A})$ denote the dominant eigenvalue of the matrix $\mathbf{A}$.

PROPOSITION 6.2 *(same as Proposition 1 in [J. Zander, 1992]) The following inequality*

$$(\mathbf{I} - \gamma_0 \mathbf{A})P \geq 0 \tag{6.6}$$

has solutions in non-negative $\mathbf{P}$ *if and only if*

$$\gamma_0 \leq \frac{1}{\rho(\mathbf{A})} = \gamma^*. \tag{6.7}$$

The power vector that satisfies the equality in (6.6) and achieves the largest SIR γ^* *is the eigenvector* $\mathbf{P}^*$ *that corresponds to the eigenvalue* $\rho(\mathbf{A})$.

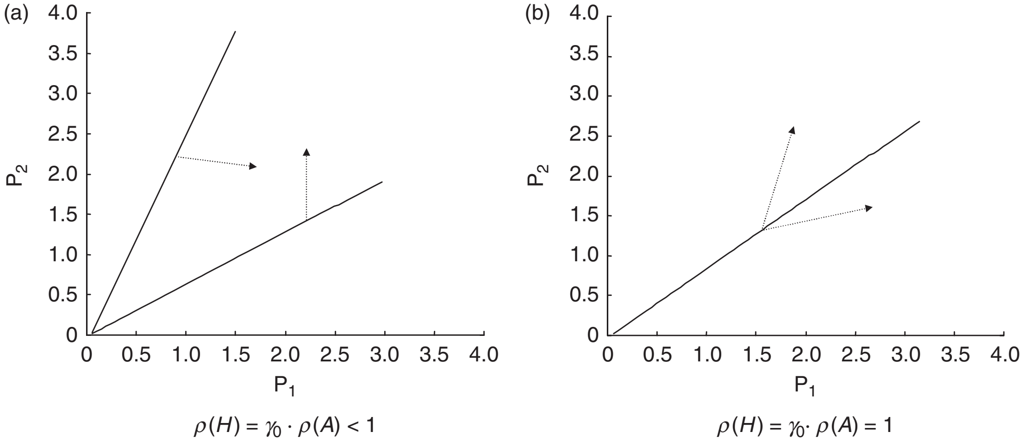
Proof We can see that

$$\rho(\mathbf{H}) = \gamma_0 \cdot \rho(\mathbf{A}).$$

Based on Proposition 6.1,

$$\rho(\mathbf{H}) = \gamma_0 \cdot \rho(\mathbf{A}) \leq 1.$$

So we have (6.7), which is also illustrated in Figure 6.3. Obviously the eigenvector $\mathbf{P}^*$ corresponding to the eigenvalue $\rho(\mathbf{A})$ is the solution to $(\mathbf{I} - \gamma^* \mathbf{A})\mathbf{P}^* = 0$ because $\mathbf{AP}^* = \rho(\mathbf{A})\mathbf{P}^*$ and $\gamma^* = \frac{1}{\rho(\mathbf{A})}$. □



$\rho(H) = \gamma_0 \cdot \rho(A) < 1$ $\rho(H) = \gamma_0 \cdot \rho(A) = 1$

Figure 6.3 Two-link case.

If $\mathbf{P}^*$ makes every user achieve the SIR γ^*, then so will the power vector $c\mathbf{P}^*$ for any scalar $c > 0$.

With positive thermal noise, based on Proposition 6.1,

$$\rho(\mathbf{H}) = \gamma_0 \cdot \rho(\mathbf{A}) < 1.$$

And $\gamma^* = \frac{1}{\rho(\mathbf{A})}$ is an upper bound on the achievable SINR level.

Example 6.2: SIR balancing for a linearly deployed system

Consider the uplink of the system illustrated in Figure 5.2. $R_1 = 1$ km, $R_2 = 2$ km and $D_{12} = 5$ km. The propagation loss exponent is 4. The thermal noise is neglected.

a) Determine the largest minimum SIR achievable in the two access ports and the corresponding transmitter powers.
b) Determine the SIR when a constant transmitter power is used at the terminals.
c) Assume the terminals apply a channel-inversion power control mechanism so that a constant received power can be reached at their respective access ports. What are the corresponding SIRs?

Solution:

a) The link gain matrix $\mathbf{G}$ is

$$\mathbf{G} = \begin{pmatrix} 1^4 & 1/3^4 \\ 1/4^4 & 1/2^4 \end{pmatrix}.$$

Correspondingly, matrix $\mathbf{A}$ is given by

$$\mathbf{A} = \begin{pmatrix} 0 & 1/3^4 \\ 1/2^4 & 0 \end{pmatrix}.$$

Solving the characteristic equation of $\mathbf{A}$ yields

$$\lambda^2 - 1/6^4 = 0,$$

so we have

$$\lambda = \pm 1/6^2.$$

The dominant eigenvalue is the larger of these two values. According to Proposition 6.2, we have

$$\Gamma_1 = \Gamma_2 = \gamma^* = 6^2 \approx 15.6 \text{ dB}.$$

The power vector $\mathbf{P}^*$ is the eigenvector corresponding to $\rho(A)$:

$$\mathbf{P}^* = c(1, 2.25).$$

b) Assume $\mathbf{P} = (1,\ 1)$. Based on (6.1),

$$\Gamma_1 \approx 81 \ (19 \text{ dB}), \quad \Gamma_2 \approx 16 \ (12 \text{ dB}).$$

The minimum SIR is only 12 dB.

c) Choose the following power vector:

$$\mathbf{P} = c\left(\frac{1}{g_{11}},\ \frac{1}{g_{22}}\right) = c'\left(1,\ \frac{g_{11}}{g_{22}}\right) = c'\left(1,\ 2^4\right).$$

The SIR can be computed as

$$\Gamma_1 \approx 5\ (7\ \text{dB}). \quad \Gamma_2 \approx 256\ (24\ \text{dB}).$$

The minimum SIR is even less than the one in case b). The channel-inversion power control may thus make the SIR worse than using no power control at all.

6.3.2 Admission control

All links will have acceptable performance if the balanced SIR γ^* is larger than the target γ_0. However, if γ^* is smaller than γ_0, the SIR balancing power control would be disastrous, as illustrated in Figure 6.4. All links would have a SIR below the threshold and the connections would be useless, indicating more connections than can be supported by the channels and an overloaded network. In this case some terminals should be removed or new connection requests should be rejected.

With the optimum power control that maximizes the number of supported links, it can be proven that some users should be completely turned off while all the remaining users have a balanced SIR. This is a scheduling issue and the links that should be turned off will depend on the scheduling algorithm implemented in the system. For instance, a stepwise removal algorithm is proposed in [J. Zander, 1992], which removes one user at a time until the required SIR is achieved in the remaining links.

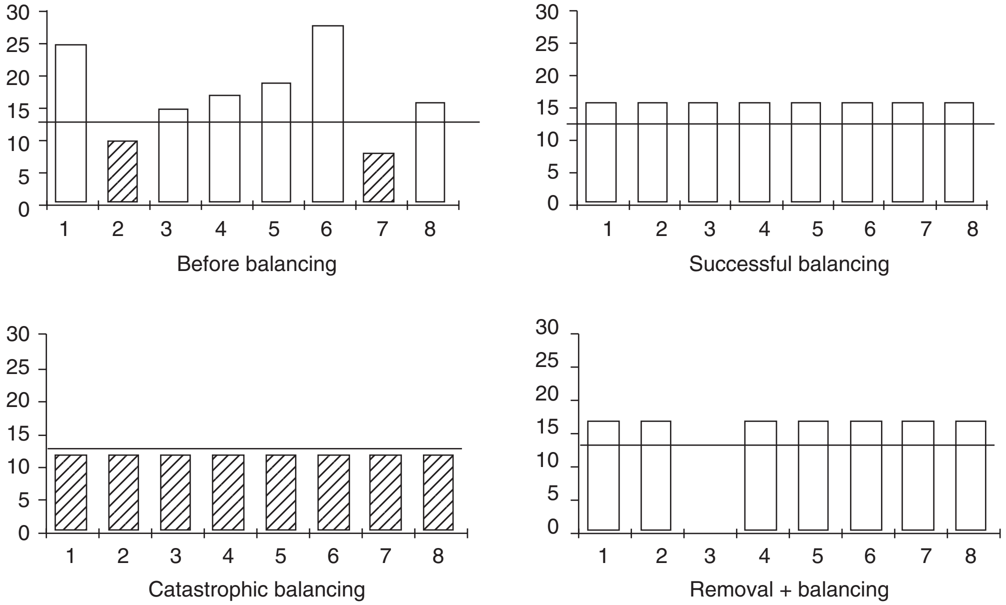


Figure 6.4 SIR balancing; eight-user case.

6.4 Distributed power control

In the previous section, we have assumed a known link gain matrix **G**, which is not feasible in practical systems because a centralized measurement and report mechanism is needed. This mechanism requires a lot of signaling overhead in the air interface. In this section, we discuss distributed power control, which does not need a centralized control and each user in the network decides its own transmission power based on local measurement. At any time instant n, the power of transmitter i is determined by

$$P_i^{(n)} = \mathbb{I}_i(\text{local measurement}), \tag{6.8}$$

where $\mathbb{I}_i$ is the power control algorithm implemented at transmitter i.

6.4.1 Iterative power control

The objective of distributed power control is still that given in (6.4). Ideally, the minimal transmission power is used and the following equality can be applied:

$$(\mathbf{I} - \mathbf{H})\mathbf{P} = \eta. \tag{6.9}$$

In Sections 6.4.1 to 6.4.4, we have assumed negligible receiver noise and that a unique and non-negative power vector $\mathbf{P}^*$ exists and solves problem (6.9). That is, $\rho(\mathbf{H}) < 1$, $(\mathbf{I} - \mathbf{H})$ is non-singular, and

$$\mathbf{P}^* = (\mathbf{I} - \mathbf{H})^{-1}\eta \geq 0.$$

In practice, it is hard to obtain each element in **H** and numerical methods, e.g., *Gaussian elimination*, cannot be applied to solve (6.9). In the following we introduce iterative methods that require only local measurement and signaling.

To solve (6.9), consider the following approach:

$$\mathbf{P}^{(n+1)} = \mathbf{M}^{-1}\mathbf{N}\mathbf{P}^{(n)} + \mathbf{M}^{-1}\eta, \quad n = 0, 1, \ldots \tag{6.10}$$

in which **M** and **N** are appropriate matrices such that

$$\mathbf{P}^* = \mathbf{M}^{-1}\mathbf{N}\mathbf{P}^* + \mathbf{M}^{-1}\eta.$$

Here vector $\mathbf{P}^{(n)}$ is the power values at iteration n. By choosing proper **M** and **N**, this iterative method will converge to the optimal power vector,

$$\lim_{n\to\infty} \mathbf{P}^{(n)} = \mathbf{P}^* = (\mathbf{I} - \mathbf{H})^{-1}\eta.$$

Define $\mathbf{M} = \mathbf{I}$ and $\mathbf{N} = \mathbf{H}$. We can obtain the following power control algorithm based on (6.10):

$$\mathbf{P}^{(n+1)} = \mathbf{H}\mathbf{P}^{(n)} + \eta, \quad n = 0, 1, \ldots$$

For transmitter i, the equivalent algorithm is

$$P_i^{(n+1)} = \frac{\gamma_0}{g_{ii}}\left(\sum_{j=1, j\neq i}^{Q} g_{ij}P_j^{(n)} + n_i\right) = \frac{\gamma_0}{\gamma_i^{(n)}}P_i^{(n)}, \quad n = 0, 1, \ldots \tag{6.11}$$

where $\gamma_i^{(n)}$ is the received SINR and $P_i^{(n)}$ the transmission power of transmitter i at iteration n. We call this algorithm distributed power control (DPC), which requires only local measurement of SINR at the receiver. In each iteration DPC allocates just the right amount of power to meet the target SINR based on the interference measurement in the previous iteration. DPC was proposed in [G. J. Foschini and Z. Miljanic, 1993] and is the same as the Jacobi relaxation method in numerical linear algebra [R. S. Varga, 1962].

6.4.2 Convergence

In this section we study the convergence properties of the iterative approach in (6.10). Define $\alpha_1, \alpha_2, \ldots$ to be the eigenvalues of $\mathbf{M}^{-1}\mathbf{N}$ and let

$$\rho(\mathbf{M}^{-1}\mathbf{N}) = \max_k |\alpha_k|.$$

Define the error vector as

$$\varepsilon^{(n)} = \mathbf{P}^{(n)} - \mathbf{P}^*.$$

Based on (6.10), $\varepsilon^{(n)}$ can be written as

$$\varepsilon^{(n)} = \mathbf{M}^{-1}\mathbf{N} \cdot \varepsilon^{(n-1)} = \cdots = (\mathbf{M}^{-1}\mathbf{N})^n \cdot \varepsilon^{(0)}. \tag{6.12}$$

The convergence property is summarized in the following proposition.

PROPOSITION 6.3 *[R. S. Varga, 1962] Given an achievable target SINR γ_0, the error vector $\varepsilon^{(n)}$ converges to the zero vector starting with any initial error vector $\varepsilon^{(0)}$ if and only if $\rho(\mathbf{M}^{-1}\mathbf{N}) < 1$.*

Observing (6.12), the convergence is at a geometric rate if it converges. In DPC, $\rho(\mathbf{M}^{-1}\mathbf{N}) = \rho(\mathbf{H})$. If the target SINR is achievable and and the thermal noise is positive, $\rho(\mathbf{H}) < 1$. So DPC will converge to $\mathbf{P}^*$.

The convergence speed of distributed power control is another important criterion in practice. In this chapter we have assumed that the link gain matrix is static and will not vary in the power control process. In real communication systems, the gain matrix may be changing continuously because of the varying wireless environment. A good power control algorithm should converge to the optimal result as quickly as possible. According to Theorem 3.2 in [R. S. Varga, 1962], a smaller $\rho(\mathbf{M}^{-1}\mathbf{N})$ will enable faster convergence. So it is important to find $\mathbf{M}$ and $\mathbf{N}$ that make $\rho(\mathbf{M}^{-1}\mathbf{N}) < 1$ as small as possible. Meanwhile these matrices should lead to the distributed form as in the DPC in (6.11).

6.4.3 General sufficient conditions for convergence

The sufficient conditions for the convergence of distributed and iterative power control algorithms are discussed in [R. Yates, 1995] and [G. W. Miao et al., 2011; G. W. Miao and G. C. Song, 2014] from different perspectives.

In [R. Yates, 1995], the so-called standard interference functions are introduced and the algorithms for power control on a single channel c_0 that have an iterative nature are described by the following general function:

$$\mathbf{P}^{(n+1)} = \mathbb{I}\left(\mathbf{P}^{(n)}\right), \tag{6.13}$$

where function $\mathbb{I} = (\mathbb{I}_1, \mathbb{I}_2, \ldots, \mathbb{I}_Q)^T$ is called the *interference function* and determines the power vector of the next iteration. The ith element of $\mathbb{I}$ indicates the effective interference power from other users that transmitter i must overcome by allocating transmission power.

DEFINITION 6.3 *With positive thermal noise, an interference function $\mathbb{I}$ is standard if, for all non-negative power vectors:*
Positivity : $\mathbb{I}(\mathbf{P}) > 0$;
Monotonicity : $\mathbf{P} \geq \mathbf{P}' \Rightarrow \mathbb{I}(\mathbf{P}) \geq \mathbb{I}\left(\mathbf{P}'\right)$;
Scalability : $\forall\alpha > 1,\ \alpha \cdot \mathbb{I}(\mathbf{P}) > \mathbb{I}(\alpha\mathbf{P})$.

The positivity is obvious. The monotonicity says that if other transmitters increase their power values, the power necessary to overcome the increased interference should be increased accordingly. With the scalability property, if $\mathbf{P} = \mathbb{I}(\mathbf{P})$, then $\alpha \cdot \mathbf{P} = \alpha \cdot \mathbb{I}(\mathbf{P}) > \mathbb{I}(\alpha\mathbf{P})$. If all the links in a network have acceptable connections under $\mathbf{P}$, the connections will be more than acceptable if the powers of all transmitters are scaled up uniformly.

For power control based on the standard interference functions, the following convergence property can be proved.

PROPOSITION 6.4 *(Theorem 2 in [R. Yates, 1995]) If the system (6.9) has a unique non-negative solution $\mathbf{P}^*$, a distributed power control algorithm based on a standard interference function will have a sequence of power vectors that converges to the solution of (6.9), starting with any non-negative power vector.*

The interference function in DPC is standard. In DPC, the power is updated according to

$$P_i^{(n+1)} = \frac{\gamma_0}{g_{ii}}\left(\sum_{j=1, j\neq i}^{Q} g_{ij}P_j^{(n)} + n_i\right) = \frac{\gamma_0}{\gamma_i^{(n)}}P^{(n)}.$$

The positivity and monotonicity are obviously satisfied. Because

$$\alpha P_i^{(n+1)} = \alpha\frac{\gamma_0}{g_{ii}}\left(\sum_{j=1, j\neq i}^{Q} g_{ij}P_j^{(n)} + n_i\right) > \frac{\gamma_0}{g_{ii}}\left(\alpha\sum_{j=1, j\neq i}^{Q} g_{ij}P_j^{(n)} + n_i\right),$$

DPC is also scalable. So it converges to the optimal solution.

Proposition 6.4 gives a sufficient condition for convergent power control. There are convergent power control algorithms that do not use standard interference functions. For instance, the following power control algorithm, called Unconstrained Second-Order

Power Control (USOPC) [R. Jäntti and S.-L. Kim, 2000] does not fulfill the conditions,

$$P_i^{(n+1)} = \omega \frac{\gamma_0}{\gamma_i^{(n)}} P_i^{(n)} + (1-\omega) P_i^{(n-1)}, n = 1, 2, \ldots$$

where $P_i^{(1)} = \frac{\gamma_i}{\gamma_i^{(1)}} P_i^{(0)}$ and $\omega = 1.1291$. In particular, the first condition is not satisfied when $\frac{\gamma_0}{\gamma_i^{(n)}} P_i^{(n)} < P_i^{(n-1)}$. However, the convergence can be proved using Proposition 6.3.

Convergence in sufficiently weak interference channels

The convergence of iterative algorithms for simultaneous power control over multiple channels are given in [G.W. Miao et al., 2011, 2014]. In the following, we introduce the single-channel version. It is shown that when the interference channels are sufficiently weak, the convergence to the optimal power vector can be guaranteed in distributed power control algorithms.

It is proved in [G. W. Miao et al., 2011] that a power control algorithm will converge to a unique optimal solution on condition that the power allocation of any transmitter is altered by a lesser amount when the other transmitters change their transmission powers by some amount. The power allocated by a transmitter and the powers of the other transmitters are related through the interference channel gains. The interference channel gains determine the variation of the reallocated power when the other transmitters change their powers. Higher interference channel gains result in stronger correlation and vice versa. When the interference channel gains are sufficiently weak, the correlation will be small and the convergence to a unique optimal solution can be guaranteed. Define the cumulative interference at transmitter i to be

$$I_i = \sum_{j=1, j \neq i}^{Q} g_{ij} P_j.$$

The following proposition explicitly shows the impact of interference channel gains on the convergence property.

PROPOSITION 6.5 *(Theorem 17.9 in [G. W. Miao and G. C. Song, 2014]) If for any transmitter i,*

$$\sum_{j=1, j \neq i}^{Q} g_{ij}^2 < \frac{1}{\sup_{I_i} \left(\frac{\partial \mathbb{I}_i}{\partial I_i} \right)^2}, \tag{6.14}$$

where $\sup_{I_i}$ *is the supremum of all non-negative* I_i*, the distributed power control algorithm in (6.13) will converge to a unique power vector* $\mathbf{P}^*$*. If the interference function* $\mathbb{I}_i$ *is linear in* I_i*, the sufficient condition (6.14) is equivalent to*

$$\sum_{j=1, j \neq i}^{Q} g_{ij}^2 < \frac{1}{\left(\frac{\partial \mathbb{I}_i}{\partial I_i} \right)^2}. \tag{6.15}$$

The left-hand side of (6.14) depends on interference channel gains only, while the right-hand side is independent of interference channel gains. Hence, interference channel gains directly impact whether a distributed power control algorithm can converge to a unique optimal solution or not. Consider an example where different users are sufficiently far away and all interference channel gains are close to zero. It is easy to see that the transmission powers of other users have almost no effect on the power allocation of the user being observed and the convergence is obvious.

Like the standard interference function approach, Proposition 6.5 gives a sufficient condition to a convergent power control. There are convergent algorithms that do not satisfy Proposition 6.5, e.g. DPC. With DPC,

$$\frac{1}{\sup_{I_i}\left(\frac{\partial \mathbb{I}_i}{\partial I_i}\right)^2} = \frac{g_{ii}^2}{\gamma_0^2}.$$

The inequality (6.14) holds only for some channel conditions. However, DPC always converges to the optimal solution, as already proved using either Proposition 6.3 or 6.4. Proposition 6.5 can be used to determine the convergence of other distributed power control algorithms.

6.4.4 Distributed power control with power constraints

So far we have assumed that the transmission power can be any value, which is not realistic. In implementation, the maximum output power is limited by the capability of power amplifiers or government regulation. This constraint is even more critical for battery-driven terminals as high transmission power will exhaust battery storage quickly. This limitation in power control can be considered by introducing a constraint,

$$0 \ \le \mathbf{P} \le \hat{\mathbf{P}}, \tag{6.16}$$

where $\hat{\mathbf{P}} = (\hat{P}_1, \hat{P}_2, \cdot\cdot\cdot, \hat{P}_Q)^T$ gives the maximum transmission powers of the Q transmitters. With the power limit, the algorithms should be modified to restrict the power values within the power constraint. For example, DPC in Section 6.4.1 can be revised as:

$$P_i^{(n+1)} = \min\{\frac{\gamma_o}{\gamma_i^{(n)}} P_i^{(n)}, \hat{P}_i\}, \quad n = 0, 1, \ldots$$

This is called *Distributed Constrained Power Control* (DCPC) [S. A. Grandhi et al., 1995]. With DCPC, the maximum power may be reached if a link has low channel quality. Maximum transmission power will lead to high power consumption and severe interference hitting other links. However, the link quality may still not be sufficiently good. Therefore, a more general scheme assuming each link has its own SINR target is given by

$$P_i^{(n+1)} = \begin{cases} \frac{\gamma_i}{\gamma_i^{(n)}} P_i^{(n)} \text{ if } \frac{\gamma_i}{\gamma_i^{(n)}} P_i^{(n)} \le \hat{P}_i \\ \check{P}_i \quad \text{if } \frac{\gamma_i}{\gamma_i^{(n)}} P_i^{(n)} > \hat{P}_i \end{cases} \tag{6.17}$$

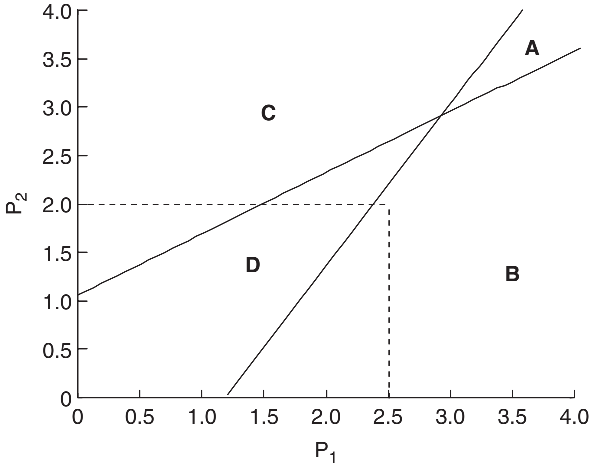


Figure 6.5 Two-user example: constrained power case.

where $0 \leq \check{P}_i \leq \hat{P}_i$. With this approach, a smaller amount of power $\check{P}_i$, instead of the maximum power, will be used when the channel quality is poor. In some situations, the power may be lowered to the minimum and the user will stay connected on the same channel and resume data transmission when the channel becomes favorable. DCPC uses a standard interference function, while the general constrained power control algorithm in (6.17) does not satisfy the second condition of the standard interference function. However, it is also a convergent power control scheme [F. Berggren et al., 2001].

6.4.5 Admission control

Assume there are power constraints in Example 6.1 and the maximum transmission powers of the terminals are 2.5 W and 2 W respectively, as illustrated in Figure 6.5. The minimum power vector that supports both terminals is $(2.96, 2.91)$. It is not possible to support both users at the same time. DCPC will converge to the power vector $(2.5, 2)$, and neither terminal is supported.

So far in this section, we have assumed the given target SIR is achievable. Like the discussion in Section 6.3.2, the distributed power control algorithms described in this section will behave undesirably if the network is overloaded, i.e. the target is not achievable within the feasible power range. In the congested case, some users should be removed in admission control so that the remaining users can have acceptable link quality. The removed users can be handed over to another channel or disconnected.

6.4.6 Dynamics of power control

So far we have assumed that the link gain matrix is fixed during power control. In this section, we discuss the impact of time-varying channels. For this purpose, consider DCPC as described in Section 6.4.4. Figure 6.6 illustrates the received SIR of a mobile terminal at a base station. The target SIR is 10 dB and DCPC power update is done

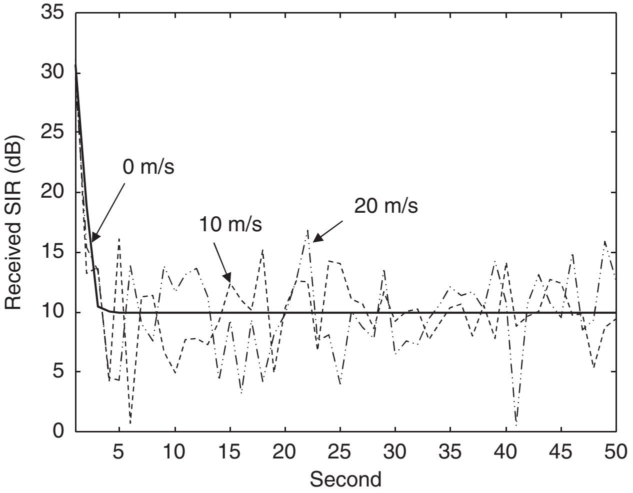


Figure 6.6 SIR of a moving terminal (DCPC power update).

every second. The terminal is moving with mean speeds, 0, 10, 20 m/s, respectively. As we can see from the figure, the target SIR 10 dB is reached in less than 5 seconds for the stationary case. With the increase in the terminal speed, the received SIR fluctuates around the target, indicating the link gain matrix changes before DCPC converges to the target SIR. The SIR fluctuations can be reduced by decreasing the power update interval, which adds a burden of frequent measurement and signaling to the system, or by developing faster power control algorithms.

6.5 Power control for elastic traffic

We have assumed the same target SINR for all users so far in this chapter. All the propositions, except SIR balancing, can be extended to the case where each user has its own target SINR. In mobile data systems, usually each user has a target SINR, as each user accesses a different type of service that requests a different transmission rate and error requirement. The transmission rate and error are closely related to the SINR, which is determined by the power control. The possibility of varying transmission rates of different users raises the problem of how to control their spectrum use most efficiently. In the following, we will introduce some fundamental insights into the issue.

6.5.1 Achievable region

Assume the transmission power of transmitter i, P_i, and the link data rate r_i can be controlled. With link adaptation, the data rate is given by

$$r_i \leq f(\gamma_i),$$

where γ_i is the achieved SIR at the receiver i. f is a monotonically increasing and strictly concave modem-dependent function that is determined by the waveform set/coding. The SINR is given by

$$\gamma_i(\mathbf{P}) = \frac{g_{ii}P_i}{\sum\limits_{j=1, j\neq i}^{Q} g_{ij}P_j + n_i}. \tag{6.18}$$

The maximum transmission powers are given by $\hat{\mathbf{P}} = (\hat{P}_1, \hat{P}_2, \cdots, \hat{P}_Q)^T$.

DEFINITION 6.4 *A data rate vector* $\mathbf{R}(\hat{\mathbf{P}}) = (r_2, \cdots, r_Q)^T$ *is achievable instantaneously if there exists a non-negative power vector*

$$\begin{aligned}&\mathbf{P} = (P_1, P_2, \cdots, P_Q) \leq \hat{\mathbf{P}}\\ &\text{such that}\\ &r_i \leq f(\gamma_i(\mathbf{P})),\ \forall i.\end{aligned}$$

DEFINITION 6.5 *A data rate vector* $\mathbf{R}^*(\hat{\mathbf{P}}) = (r_1^*, r_2^*, \ldots, r_Q^*)$ *is achievable in the average sense if*

$$\begin{aligned}&\mathbf{R}^* = \sum_k \alpha_k \mathbf{R}_k\\ &\text{where}\\ &\alpha_k \in [0,1],\ \sum_k \alpha_k = 1,\end{aligned}$$

where all the $\mathbf{R}_k$ *are instantaneously achievable data rate vectors.*

If a set of data rate vectors $\mathbf{R}_k$ are achievable instantaneously, they can be switched in time. For example, each of them can be used during some fraction of the time α_k, which is TDMA. This will yield the average rate

$$\mathbf{R}^*(\hat{\mathbf{P}}) = (r_1^*, r_2^*, \ldots, r_Q^*)^T.$$

Assume link i has a minimum data rate requirement, $r_{i,\min}$, and any excess data rate is consumed by the user as well. The network has an interest in providing as high a data rate as possible. Then we can consider the following optimization problem:

$$\max \sum_i r_i^*(\hat{\mathbf{P}}) \tag{6.19}$$

subject to

$$r_i^*(\hat{\mathbf{P}}) \geq r_{i,\min}, \quad \forall i.$$

Below we give two sample relations between the data rate and the SINR γ:

Case A: $f(\gamma) = c\gamma$

Case B: $f(\gamma) = c' \log(1+\gamma)$

Case B is the Shannon limit for a band limited channel. Case A corresponds to the low SINR regime and can be seen as a special example of case B. In both cases there is a one-to-one mapping between the instantaneously achievable data rate and the SINR.

To simplify the analysis, consider a two-user case and simplify (6.18) to

$$\begin{array}{ccc} P_1 & -\gamma_1 a_{12} P_2 & \geq \gamma_1 \eta_1 \\ -\gamma_2 a_{21} P_1 & +P_2 & \geq \gamma_2 \eta_2 \end{array}$$

where

$$a_{ij} = \frac{g_{ij}}{g_{ii}} \qquad \text{and} \qquad \eta_i = \frac{n_i}{g_{ii}}$$

and γ_i is the minimum SIR to achieve the instantaneous data rate r_i. Solving for P_1 and P_2, we have

$$\begin{array}{l} 0 \leq P_1 = \frac{\gamma_1(\eta_1 + \gamma_2 a_{12} \eta_2)}{1 - \gamma_1 \gamma_2 a_{12} a_{21}} \leq \hat{P}_1 \\ 0 \leq P_2 = \frac{\gamma_2(\eta_2 + \gamma_1 a_{21} \eta_1)}{1 - \gamma_1 \gamma_2 a_{12} a_{21}} \leq \hat{P}_2. \end{array} \tag{6.20}$$

In addition, both power values should be non-negative and we have the following necessary condition:

$$\begin{array}{l} 1 - \gamma_1 \gamma_2 a_{12} a_{21} > 0 \\ \gamma_1 \gamma_2 < \frac{1}{a_{12} a_{21}}. \end{array}$$

Now apply cases A and B to (6.19); the results are illustrated in Figures 6.7 and 6.8 respectively. We can see that the achievable instantaneous data rate regions, those below the solid lines, are not convex, while the average data rate regions that are below the dotted lines are. The average data rate region is given by a set of linear equations and the boundaries are determined by the straight lines. Figure 6.7 also illustrates a minimum rate requirement $r_i > r_{i,\min}$. When the maximal power is 5, there is no rate pair that meets the constraints, neither instantaneously nor in the average sense. When the maximal power is 10, the data rate requirements can be satisfied in the average sense but not instantaneously. With maximal power 15, the constraints can be met in both senses and the sum rate is maximized by time multiplexing. The maximal average sum rate is on one of the corners of the achievable data rate region, i.e. at either point Q_1 or Q_2, and Q_1 is the optimum in this example.

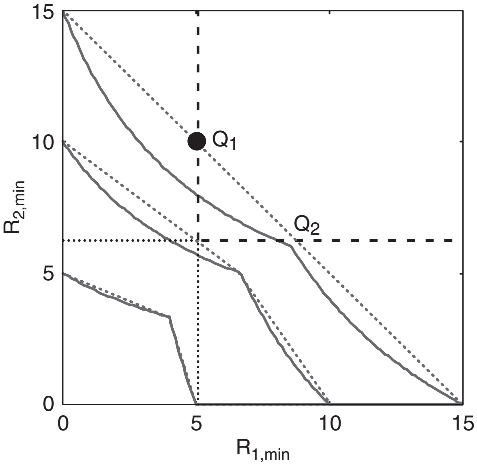


Figure 6.7 Achievable rates for various maximal transmitter powers. No bandwidth limitation (Case A). Relative maximal powers 5, 10 and 15. $a_{12} = 0.05$, $a_{21} = 0.02$, $\eta_1 = \eta_2 = 1$.

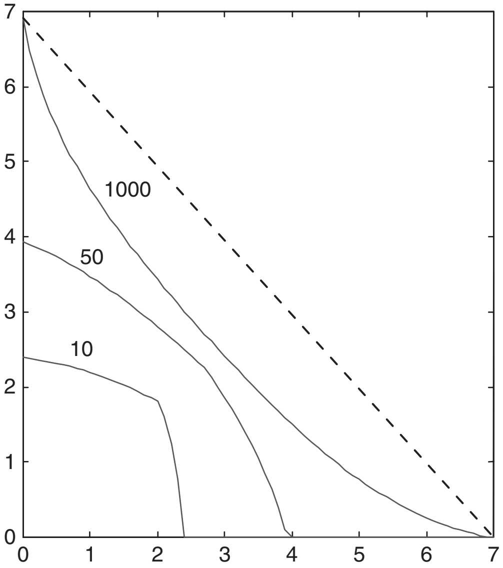


Figure 6.8 Achievable rates for various maximal transmitter powers. Band-limited (Case B). Relative maximal powers 10, 50 and 1000. $a_{12} = 0.05,\ a_{21} = 0.02,\ \eta_1 = \eta_2 = 1,$ c' = 1.

In case B, the instantaneous rate regions are concave for low maximal transmitter powers. The regions become convex for higher transmission powers. Higher average sum data rates can be achieved by time-multiplexing the transmissions of the two users.

6.5.2 Distributed power control for wireless data

In this section, we discuss power control algorithms for data applications; the general data rate and SIR relationship of case B will be used. That is,

$$r_i(P_i) = c' \log(1+\gamma_i) = c' \log\left(1 + \frac{g_{ii}P_i}{\sum\limits_{j=1, j\neq i}^{Q} g_{ij}P_j + n_i}\right).$$

The objective is to control the power values to maximize the network throughput, and the optimization problem is

$$\max_{\{P_i\}} \sum_i r_i(P_i) = \sum_i c' \log\left(1 + \frac{g_{ii}P_i}{\sum\limits_{j=1, j\neq i}^{Q} g_{ij}P_j + n_i}\right) \tag{6.21}$$

subject to

$$r_i \geq r_{i,\min}, \quad \forall i.$$

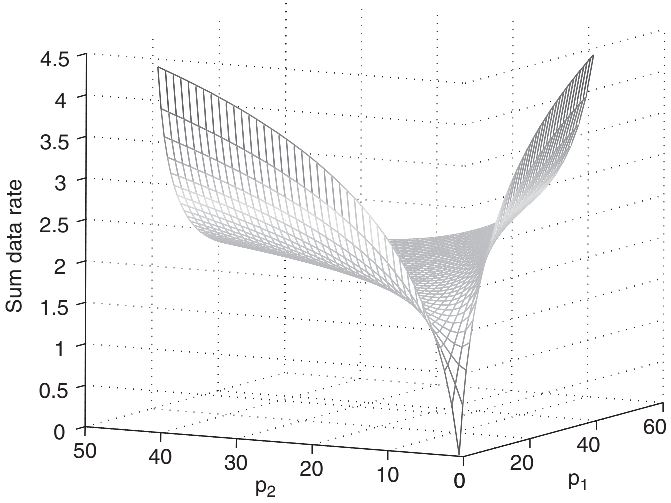


Figure 6.9 Sum network throughput for various transmitter power values ($a_{12} = a_{21} = 1$ and others are 1).

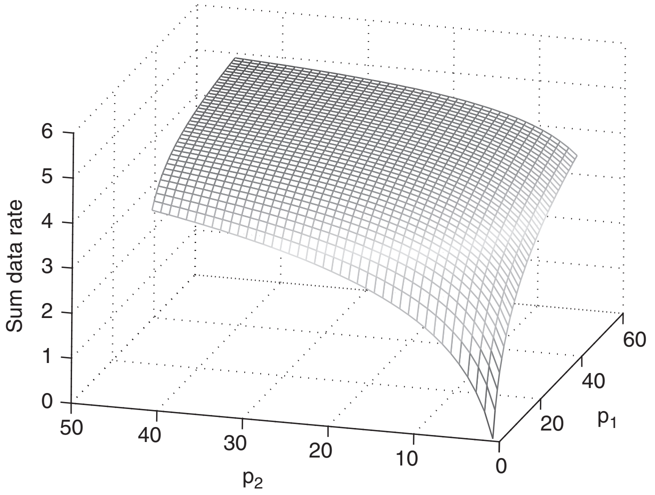


Figure 6.10 Sum network throughput for various transmitter power values ($a_{12} = a_{21} = 0.1$ and others are 1).

The objective function in (6.21) is non-concave and many local maxima may exist depending on channel conditions. To illustrate this, consider the two-user case,

$$r_1(P_1) + r_2(P_2) = c^{'} \log\left(1 + \frac{g_{11}P_1}{g_{12}P_2 + n_1}\right) + c^{'} \log\left(1 + \frac{g_{22}P_2}{g_{21}P_1 + n_2}\right).$$

Figures 6.9 and 6.10 illustrate the corresponding relationship between the sum data rate and the two power values. The results in Figure 6.9 show that when inter-user interference is the dominant source of disturbance (e.g. shorter ranges like wireless LANs), the maximal total (sum) data rates for single transmissions are higher. This holds true in many cases, but not always. When the inter-user interference is not dominant, e.g. longer ranges like inter-cell interference in cellular networks, maximal sum rates are mostly reached for simultaneous transmission schemes, as demonstrated in Figure 6.10.

Note that the characterizations are general in the sense that they are not sensitive to the choice of modulation schemes and therefore apply for CDMA type waveforms as well. As a practical consequence, the results demonstrate that for short range, interference-limited systems (e.g. indoor, microcell), coordinated orthogonal multiplexing provides better performance than simultaneous transmissions (e.g. CDMA). In Chapters 3 and 4, the MAC and scheduling protocols and algorithms are designed to coordinate the orthogonal transmissions of all users in these interference-limited systems. For longer ranges when noise (i.e. coverage) comes into play, non-orthogonal schemes have advantages and power control algorithms are needed to determine the transmission power values.

Generally speaking, the non-concave power control problem in (6.21) is hard to solve. In [L. P. Qian et al., 2009], a centralized algorithm called MAPEL is proposed to find an optimal solution based on increasingly accurate approximation of the feasible SINR region. However, the algorithm has exponential complexity and cannot be used in practice. Furthermore, a centralized algorithm requires complete channel knowledge, including the interference channels, $\{g_{ij}\}$. In the following, we introduce distributed heuristic power control algorithms.

With distributed power control, each transmitter decides its power level based on local observations. The power of transmitter i is

$$P_i^{(n+1)} = f_i(A_i^{(n)}),$$

where $A_i^{(n)}$ is the parameter set observed locally in the previous iteration. For example, it can be the previous transmission power $P_i^{(n)}$, or the effective interference of other users plus noise that user i must overcome, $I_i^{(n)}$, or both.

For delay-sensitive traffic with hard SIR requirements, some examples of $f_i(P_i^{(n)}, I_i^{(n)})$ have been given in Section 6.3. Another example assuming each user has its own target SIR, γ_i, is to allocate a power value such that the SIR target is just met in iteration $n+1$ based on the observation in iteration n, i.e.

$$\frac{P_i^{(n+1)} g_{ii}}{I_i^{(n)}} = \gamma_i,$$

and

$$P_i^{(n+1)} = f_i(P_i^{(n)}, I_i^{(n)}) = \frac{\gamma_i}{g_{ii}/I_i^{(n)}} = \frac{\gamma_i}{\gamma_i^{(n)}} P^{(n)}. \tag{6.22}$$

This can be seen as a generalization of the DPC algorithm in Section 6.4.1. Similar to the discussion of Proposition 6.4, (6.22) is also a standard interference function. Therefore, if the SIR requirements are feasible, starting with any non-negative power values, this distributed algorithm will converge to an equilibrium where no user wants to further change its power. Furthermore, in the equilibrium, the SIR requirements of all users will be just met, as otherwise some users will change their power values.

As discussed in Chapter 4, there are many other types of services other than delay-sensitive ones. For these delay-tolerant services, e.g. email, web browsing, and file transfer, it is not desirable to use the power control algorithms that are developed for supporting strict SIR requirements. These delay-tolerant services usually desire a high data rate and, as shown in Figure 4.11, higher data rates will always result in higher utility (user satisfaction). One simple approach is for each user to allocate power to selfishly maximize its own data rate, i.e.

$$P_i^{(n+1)} = \arg\max_P r_i^{(n+1)} = \arg\max_P c' \log\left(1 + \frac{g_{ii}P_i}{\sum_{j=1,j\neq i}^{Q} g_{ij}P_j + n_i}\right).$$

This may not work because all users will simply use the highest power possible and the network performance might be poor because of excessive interference. The problem can be resolved by introducing the concept of pricing. The pricing mechanism introduces a cost for each user to use a certain amount of power and thus regulates aggressive power control behavior to reduce interference. A linear pricing example is

$$U_i = r_i^{(n+1)} - \mu_i P = c' \log\left(1 + \frac{g_{ii}P_i}{\sum\limits_{j=1,j\neq i}^{Q} g_{ij}P_j + n_i}\right) - \mu_i P,$$

where μ_i is the pricing parameter and

$$P_i^{(n+1)} = \arg\max_P U_i.$$

Some examples of the utility function, U_i, are illustrated in Figure 6.11. We can see that different pricing parameters will result in different power values and using a proper μ_i will help control network interference and improve network throughput.

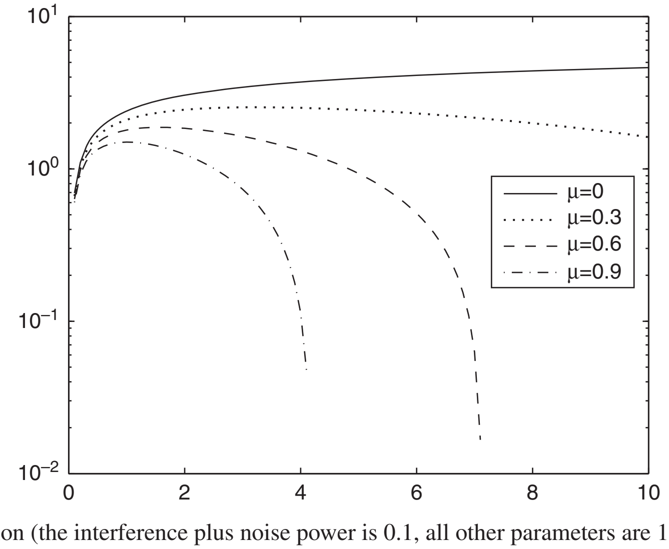


Figure 6.11 Utility function (the interference plus noise power is 0.1, all other parameters are 1).

U_i is concave in P_i and its maximum can be found by setting its first-order derivative to be zero as follows,

$$\frac{dU_i}{dP_i} = \frac{c'}{\frac{I_i}{g_{ii}} + P_i} - \mu_i = 0,$$

where $I_i = \sum_{j=1, j \neq i}^{Q} g_{ij} P_j + n_i$. Based on the above equation, we have the power control formula

$$P_i^{(n+1)} = \max\left(\frac{c'}{\mu_i} - \frac{I_i^{(n)}}{g_{ii}}, 0\right). \tag{6.23}$$

This power control function is not a standard interference function and there might be multiple equilibria. Proposition 6.5 can be used to decide when the power control will converge to a unique optimal solution, which depends on how weak interference channels are. Similarly, it is verified in [T. Alpcan et al., 2002] that if the spreading gain in a CDMA system is sufficiently large, i.e. interference channels are sufficiently weak, the algorithm in (6.23) will converge to a unique equilibrium. The interested reader may refer to [T. Alpcan et al., 2002] for more thorough discussions of the global stability and convergence of the power control in (6.23).

6.6 Power control in DS-CDMA cellular systems

In direct sequence CDMA (DS-CDMA) communication systems, several users share the same radio bandwidth simultaneously, through the use of specific spreading codes. These codes are not exactly orthogonal, which creates interference between active users of the same cell and the users from other nearby cells. As we have seen in Chapter 5, this interference increases as the number of users increases, and hence the CDMA system is said to be interference-limited. Apart from this, the link quality exhibits considerable fluctuation caused by fading multipath channels, as mobile users move within the service area, and path loss effects. The total dynamic range of path loss in cellular systems is of the order of 80 dB and the fluctuations caused by fading multipath effects is of the order of 20 to 30 dB. These fluctuations cause big variations in the received signal-energy-to-noise-power-spectral-density ratio for the different users, particularly on the uplink where mobile users are transmitting from different places within the cell. In fact, if these fluctuations are not taken care of, mobile users will easily jam each other making reliable simultaneous transmission impossible. This jamming problem is known as *the near–far problem* where users near the base station jam users far away from the base station, as illustrated in Figure 6.12. To mitigate this problem and improve the capacity of DS-CDMA, power control should be used. Power control can mitigate these effects and allows better battery life by controlling the transmit power of the users such that the received power for each user at the base station is equal (assuming they each employ the same service, such as speech). Power control should also ensure that the received SIR is good enough to maintain the required quality of the service. Transmit power control is performed on both the downlink and uplink. In the downlink, the power

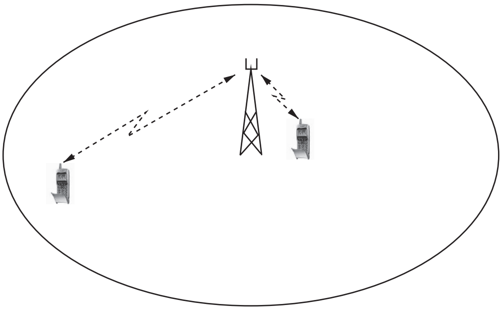


Figure 6.12 When both users are transmitting with the same power, the faraway user will be jammed by the nearby user.

control is used for cases where users experience large path losses such that the received signal is of the order of the noise, but the power control requirement is not as stringent as it is for the uplink. In the uplink, power control is used to solve the near–far problem and the effects of fading multipath channels. Since, in DS-CDMA systems, mobile users are using the whole band, we may see resource allocation in DS-CDMA as just power allocation where each user should employ the transmit power that ensures its quality of service.

For the capacity calculation of DS-CDMA in Chapter 5, we have assumed perfect transmit power control. However, transmit power control is never perfect due to the fast variations of the radio channel and user mobility. Therefore, what is important is to identify a transmit power control scheme that is fast, does not require a lot of measurements, and does not require a large feedback bandwidth.

Let us consider the discrete power control discussed in the previous sections. The transmit power in DPC for user i at time instant $n+1$ is defined by the iterative process

$$P_i^{(n+1)} = P_i^{(n)} + 10\log_{10}\left(\frac{\gamma_t}{\gamma_i}\right), \quad \text{(in dB)}, \tag{6.24}$$

which has a power update that can be very large and requires a large signaling bandwidth to deliver the exact value of $\gamma_i^{(n)}$ to the transmitter. This does not consider the presence of measurement errors.

In order to cope with these practical limitations, a slightly different approach is employed in DS-CDMA systems. In trying to reduce the feedback information, the transmit power control on the uplink consists of three parts that are running simultaneously, the open loop power control, the inner loop power control (or closed fast loop control), and the outer loop power control.

The open loop power control handles the wide dynamic range and represents the ability of the mobile user to set its output power to a specific value without the help of the base station. The main objective of open loop power control is to equalize the path loss between the mobile user and the base station. This is achieved by measuring the received signal strength from the base station and adjusting the transmit power accordingly, as illustrated in Figure 6.13. Here the mobile transmit power is set as

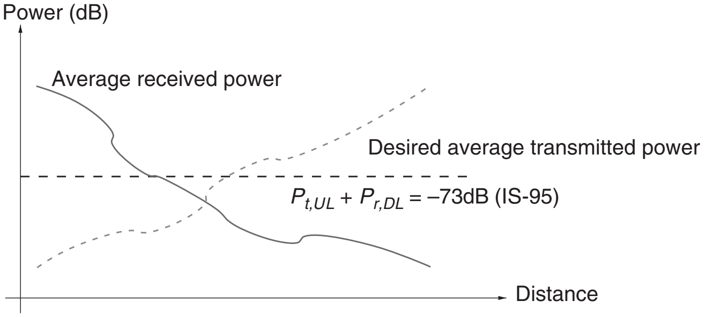


Figure 6.13 Open loop power control in DS-CDMA systems.

follows:

$$P_{t,UL} + P_{r,DL} = c_0, \quad \text{(in dB)} \tag{6.25}$$

where $P_{t,UL}$ is the mobile user transmit power, $P_{r,DL}$ is the received power from the base station, and c_0 is a constant (for IS-95 we have $c_0 = -73$ dB (mW^2)). Denoting by L_p the link path loss, the mobile user transmit power can be written as

$$P_{t,UL} = c_0 - P_{t,DL} + L_p \rightsquigarrow P_{r,UL} \approx c_0 - P_{t,DL}, \tag{6.26}$$

where $P_{t,DL}$ is the transmit power portion allocated to the mobile user, and $P_{r,UL}$ is the received power at the base station receiver.

It can be seen from (6.26) that the path loss has been equalized and with a proper selection of the constant c_0 the mobile user can bring its transmit power quite close to the required value. In the IS-95 system, this is accomplished by utilizing the automatic gain control (AGC) circuit of the receiver. The AGC circuit operate on the receiver's IF frequency amplifier so that the input to the receiver's analog/digital (A/D) converters is held constant. The AGC control is used to control the gain of the transmitter IF amplifiers exactly in step with the receiver's IF gain. Thus, if the mobile terminal moves closer to the base station, increasing the received signal level, the receiver AGC will reduce the receiver IF amplifier gain, and the transmitter IF gain. Due to the lack of reciprocity in radio channels, open loop power control is not very accurate. Studies in the literature have shown that open loop power control can reduce the dynamic range variations to about 9 dB. To reduce this inaccuracy a closed loop power control is employed on top of the open loop power control. As the name indicates, closed loop power control requires a feedback measurement to adjust the transmitter powers. To reduce the required feedback information, commercial DS-CDMA systems employ a very simple but fast closed loop power control. That is, the receiver measures the signal quality (SINR) of the transmitter and compares it with a target threshold. If the received quality is less than the target, the receiver will send a binary power-up command to the transmitter. Otherwise, a power-down command will be delivered to the transmitter. If the measurement occurs N times per second, then a power control scheme with a speed of N bps is obtained. Upon receiving the power control bits, either up or down, the transmitter will increase or decrease its power by a fixed amount, e.g., 1 dB up or

1 dB down. For instance, in the uplink of the IS-95 system, an 800 bps power control speed was suggested. That is, SIR measurement is made every 1.25 milliseconds, giving an 800 bps power control command stream in the downlink. The value represented by the closed loop power control is converted to an analog voltage and then added to the open loop control voltage (the receiver AGC signal), and applied to the transmitter gain control circuits. The effect of this control is that even in a fading channel, the received power is maintained constant so as to achieve the required block error rate (BLER) target. However, with this updating speed, the power control algorithm is still not able to completely compensate for the fast fading variations and the received SINR will experience some inaccuracies. This inaccuracy depends on the movement of the mobile units. Hence, the power controlled SINR will experience some random variations that depend on the number of interferers and the maximum Doppler frequency of the channel. It has been shown in [A. M. Viterbi et al., 1993] that the power controlled SINR of the IS-95 uplink is approximately log-normally distributed with a standard deviation between 1 and 2 dB.

Third-generation systems such as ETSI WCDMA [Standard-95 TIA/EIA Interim, 1993] have similar closed power control to the uplink of the IS-95 but with a two-times higher speed, 1600 times per second. Unlike the IS-95, this kind of power control is applied to both uplink and downlink. Note that increasing power control speed would improve the power control efficiency. However, at the same time, it will increase the measurement and signaling burden.

One problem with fast power control is the spikes experienced in the transmit power when deep fades are encountered. This may be needed to ensure the connection but it also introduces interference to neighboring cells where the mobile units may not necessarily be experiencing adverse channel conditions. To solve this problem, the rate of fast power control can be adjusted to suit the need. For example, for non-real-time services, a higher BLER can be tolerated. As a result, it is permissible to be in a fade and lose packets, leaving it to the radio link control (RLC) to retransmit. Hence, closed loop power control allows for lower rates, by which it is meant that the transmit power control (TPC) bits do not change from slot to slot. For the downlink power control, the DPC MODE controls this behavior, enabling the use of the same TPC bits for three slots. For the uplink power control, the power control algorithm tells the mobile unit how the TPC bits are processed. For the slower rate, the mobile unit considers the TPC bits from five slots before changing its power [3GPP, 2012].

In the closed loop power control, the target SIR values are controlled by another loop, called *outer loop power control*. The outer loop power control aims at providing the required quality (no worse, no better). This loop measures the link quality, typically block error rates, and adjusts the SIR targets accordingly, ensuring that the lowest possible SIR target is used at all times to maximize the network capacity and that the fast power control strategy is operating correctly. The principle of the outer loop power control is illustrated in Figure 6.14. Outer loop power control is used in the uplink of IS-95 and both uplink and downlink of WCDMA systems. A simple algorithm, the sawtooth algorithm [A. Sampath et al., 1997], is usually employed for outer loop power control. In this algorithm the SIR target is adjusted step-wise with a different step size

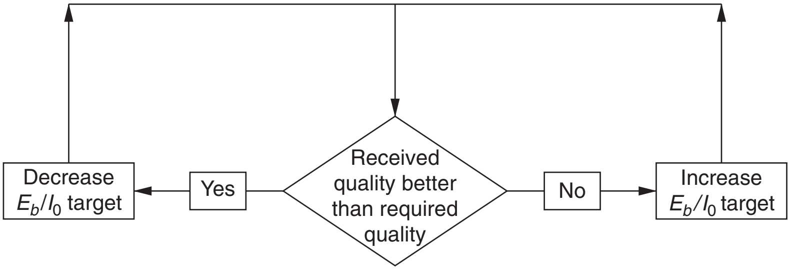


Figure 6.14 Outer loop power control procedure in DS-CDMA systems.

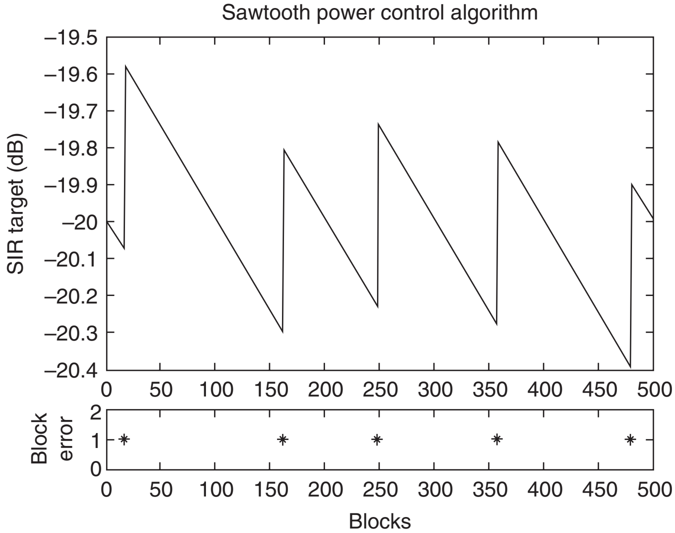


Figure 6.15 Outer loop power control procedure in DS-CDMA systems.

each for step-up and step-down. When a data block is received with no error the SIR target (in dB) is decreased by a value equal to $\mathrm{SIR}_{\text{step-down}}$ with

$$\mathrm{SIR}_{\text{step-down}} = \mathrm{SIR}_{\text{step}} \times \mathrm{BLER}_{\text{target}}. \tag{6.27}$$

On the other hand, when a data block is received with error the SIR target (in dB) is increased by a value equal to $\mathrm{SIR}_{\text{step-up}}$ with

$$\mathrm{SIR}_{\text{step-up}} = \mathrm{SIR}_{\text{step}} - \mathrm{SIR}_{\text{step-down}} \tag{6.28}$$

where $\mathrm{SIR}_{\text{step}}$ determines the convergence rate of the algorithm to converge to the ideal $\mathrm{SIR}_{\text{target}}$ and the stability of the obtained BLER.

Figure 6.15 illustrates the operation of a typical sawtooth algorithm for a BLER target of 1% and a step size of 0.5 dB [S. Yagnaraman, 2010]. It can be seen that the SIR target decreases slowly (linearly) when the received data blocks are correct but increases rapidly when a data block is received with error. We also notice the sawtooth shape of the SIR target given by this algorithm. The behavior of the sawtooth algorithm gives

Figure 6.16 Network deployment.

good protection for the required quality of service but may overestimate the required transmit powers and may take a long time to converge.

Exercises

6.1 Assume that in a wireless cellular system, each terminal controls its power to achieve a constant received power at its receiver. Estimate the reduction in average transmission power compared to a system with constant power. Assume the terminals are uniformly distributed over circular cells and the propagation loss exponent is α.

6.2 As illustrated in Figure 6.16, a cellular operator has installed a small mobile telephony system along a street. The system has only four base stations located at the following distances (coordinates), measured from one end of the street:

$$(100,200,300,400).$$

The system uses only one channel. The propagation loss is proportional to the fourth power of the distance, i.e. $P_r = P_t/d^4$. At one time, two terminals, A and B, at (150, 280), are connected. Base stations that do not serve any terminal are kept off.

a) Determine the link gain matrix in dB scale for this channel set.
b) Assume the base stations and the terminals use the same transmit power. The two terminals are connected to the base stations such that their uplink SIRs are maximized. Determine the SIRs for all base stations and terminals.
c) Observe the uplink. Assume the two terminals can have a maximum transmit power of 1 W. We require terminal A to have a minimum SIR of 20 dB. What is the maximum SIR of terminal B? What power values should the two terminals use?

6.3 A small mobile telephony system is installed along a street and the system has six base stations located at the following coordinates,

$$(100,200,300,400,500,600).$$

The channel is assigned to four mobiles at the following locations,

$$(50,280,420,545).$$

a) Assuming the path loss exponent is four, determine the link gain matrix.
b) Assume all base stations and mobile terminals use the same constant transmission power. The mobiles are connected to the base station with the lowest path loss. Base stations that do not serve any terminals are turned off. Determine the SIRs.

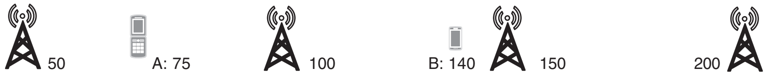


Figure 6.17 Network deployment.

c) Assume a power control scheme that maintains a constant received power at the receiver. Determine the transmission powers and the SIR in the base stations and mobiles.
d) Compare the results in c) with those using the SIR-balancing power control scheme.

6.4 A cellular operator has installed a small mobile telephony system along a street (see Figure 6.17). The system has only four base stations located at the following distances (coordinates, in meters), measured from one end of the street:

$$(50, 100, 150, 200).$$

The system uses only one channel. The propagation loss is proportional to the fourth power of the distance, i.e. $P_r = s * P_t/d^4$, where the fading s is uniformly distributed between 0.2 and 1. At one time, two terminals, A and B, at 75 and 140, are connected. Base stations that do not serve any terminal are kept off. A terminal will be in outage if its SIR is below 10 dB.

a) Determine the average link gain matrix for this channel set.
b) Determine the uplink outage probabilities of the two terminals if A is connected to the second base station and B to the fourth. $P_t = 1$ W.
c) Determine the corresponding downlink outage probabilities of the two terminals in (b).

6.5 A wireless cellular system has three cells and two channels. There are five active terminals. SIR-balancing power control is employed. Both channels may be used if an SIR of at least $\gamma_0 = 18$ dB can be achieved. The link gain matrix is given by

$$\mathbf{G} = \begin{pmatrix} 2\cdot 10^{-7} & 2\cdot 10^{-7} & 4\cdot 10^{-8} & 2\cdot 10^{-9} & 2\cdot 10^{-10} \\ 2\cdot 10^{-9} & 1\cdot 10^{-7} & 2\cdot 10^{-6} & 1\cdot 10^{-7} & 2\cdot 10^{-9} \\ 2\cdot 10^{-10} & 2\cdot 10^{-9} & 4\cdot 10^{-9} & 2\cdot 10^{-7} & 2\cdot 10^{-7} \end{pmatrix}$$

a) Is there a channel allocation such that the SIR requirement is met in the downlink for all terminals?
b) Determine the SIRs that can be achieved in the uplink for this channel allocation. What is the conclusion?

6.6 Consider two terminals connected to two access ports, one for each, on the same channel. The uplink link gain matrix is given by:

$$\mathbf{G} = \begin{pmatrix} 1/3 & 1/21 \\ 1/20 & 1/4 \end{pmatrix}$$

The receiver noise level is 0.1 W.

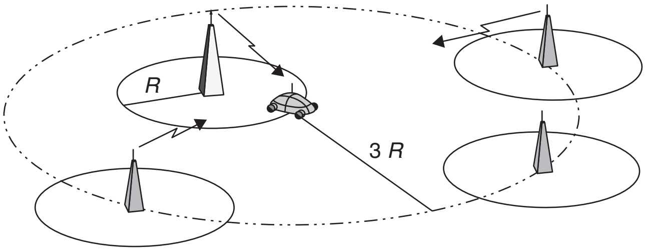


Figure 6.18 Transmission at a cell edge.

a) Is it possible to make the minimum received SIR greater than 7 dB (neglect thermal noise)?
b) Assume the target SINR is 6 dB. What minimal transmission powers should be used to achieve the target SINR in both terminals?
c) Assume the power vector is initially $P^{(0)} = (1.87, 0.52)$. How many iterations are needed for DCPC to support both mobiles with the received SINR of 5 dB? Assume the target SINR is 6 dB.
d) Assume the maximum transmission power is 3 W and answer question b) again. Can both terminals be supported in this case? If not, which one should be removed?

6.7 The path losses in two radio links using the same channel are characterized by the following gain matrix:

$$\mathbf{G} = 10^{-8} \begin{pmatrix} 1 & 0.02 \\ 0.05 & 0.3 \end{pmatrix}$$

$$\mathbf{P}_{rx,i} = \mathbf{P}_{tx,j}\mathbf{G}_{ij}$$

a) If the maximum transmit power in each of the transmitters is 1 W and the noise power is 10^{-9} W, what is the largest minimum signal SINR that can be maintained in both receivers simultaneously?
b) What happens if the maximum transmit power is 10 W? Infinite?

6.8 In a certain part of a network, with identical circular cells of cell radius R, we are considering a downlink transmission to a mobile station (in focus) at the cell edge, distance R from its home base station (see Figure 6.18). There are three potential interfering base stations at a distance of $3R$ from the mobile station (in focus). The SNR at this mobile station is 30 dB and in order to run the service at the mobile station we need at least an SINR of 18 dB. The other mobile stations are assumed to be uniformly distributed in the cells and the path loss exponent is assumed to be $\alpha = 4$. We assume that all base stations adjust their transmit power so that every mobile station gets the same received power. Let q be the probability that an interfering base station is transmitting and assume the probability is the same for the three base stations.

Figure 6.19 shows the cumulative probability distribution function of $\sum_{i=0}^{n} r_i^4$, where r_i is the user distance from the center of a circle cell normalized by the cell radius R, where the users are uniformly distributed.

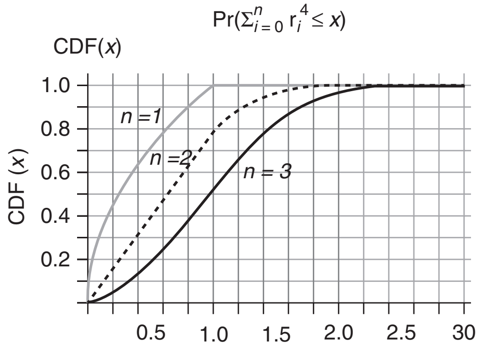


Figure 6.19 Cumulative probability distribution function.

a) Show that the service at the focused mobile station is running with 100% probability if we have at most one transmitting interferer.
b) Let q=0.5. What is the probability that the service is running while there are exactly two transmitting interferers?
c) How much activity q can we accept if we want the service of the focused mobile station to run with 95% probability?

6.9 Two radio links are operating in the TV band and are required to protect TV receivers from harmful interference. The TV set in question receives a signal level of –75 dBm and requires an SIR of at least 20 dB to avoid picture distortion. The path gains are given in Figure 6.20. The maximum (unrestricted) transmitter power of the radio transmitters T1 and T2 is 1 W and the total (external) noise level is –100 dBm. The data rate in each link is dependent on the signal to interference plus noise ratio γ and is given by

$$R = 2\log_2(1+\gamma)\text{ Mbps.}$$

a) Determine the maximal rates in link 1 and 2 if T1 and T2 transmit simultaneously using the highest possible power that will protect the TV set.
b) Compare the total data rate in a) with the data rate that can be achieved by time-sharing, letting one of the transmitters transmit at a time, 50% of the total time each.

6.10 In the wireless network shown in Figure 6.21, three transmitters T1, T2 and T3 are communicating with their respective receivers R1, R2 and R3. The path losses (in dB) between the transmitters and receivers are given in the figure. The maximum transmitter power is 1 W. The receiver noise power is –120 dBw. The data rate that can be used in each link is related to the SINR Γ in a link according to

$$R = f(\Gamma) = 50\log_2(1+\Gamma)\text{ kbit/s.}$$

a) Assume that initially transmitters T1 and T2 are transmitting (to receivers R1 and R2). What is the maximal feasible average sum rate in two links?
b) Now the link T3–R3 is activated. Is it possible to achieve 100 kbit/s in all three links?

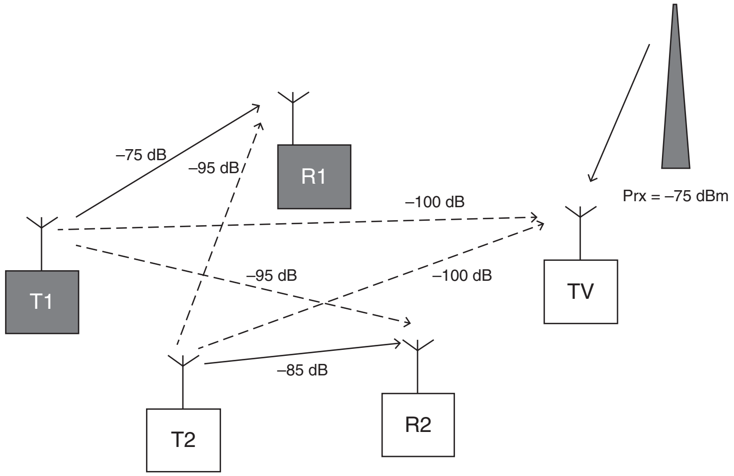


Figure 6.20 Network conditions.

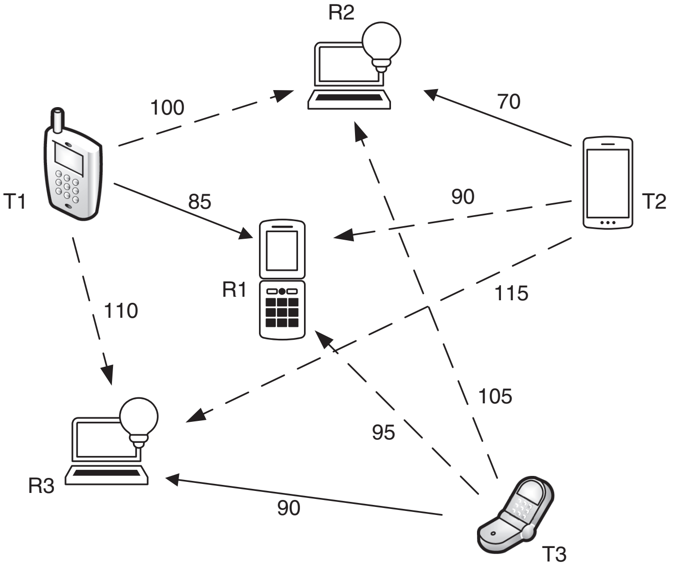


Figure 6.21 A network of three pairs of users.

References

3GPP. 2012 (March). *Physical Layer Procedures (FDD).* TS 25.214. 3rd Generation Partnership Project (3GPP).

T. Alpcan, T. Basar, R. Srikant and E. Altman. 2002. CDMA uplink power control as a noncooperative game. *Wireless Networks*, 8, 659–670.

F. Berggren, R. Jäntti and S.-L. Kim. 2001. A generalized algorithm for constrained power control with capability of temporary removal. *IEEE Trans. Veh. Tech.*, 50(6), 1604–1612.

G. J. Foschini and Z. Miljanic. 1993. A simple distributed autonomous power control algorithm and its convergence. *IEEE Trans. Veh. Tech.*, 42, 641–646.

S. A. Grandhi, J. Zander and R. Yates. 1995. Constrained power control. *Wireless Personal Communications*, 1, 257–270.

R. Jäntti and S.-L. Kim. 2000. Second-order power control with asymptotically fast convergence. *IEEE J. Sel. Areas Commun.*, 18, 447–457.

G. W. Miao, N. Himayat, G. Li and S. Talwar. 2011. Distributed interference-aware energy-efficient power optimization. *IEEE Trans. Wireless Commun.*, 10(4), 1323–1333.

G. W. Miao and G. C. Song. 2014. *Energy and Spectrum Efficient Wireless Network Design*. Cambridge: Cambridge University Press.

L. P. Qian, Y. J. Zhang and J. Huang. 2009. Mapel: Achieving global optimality for a non-convex wireless power control problem. *IEEE Trans. Wireless Commun.*, 8(3), 1553–1563.

A. Sampath, O. Sarath Kumar and J. M. Holtzman. 1997. On setting reverse link target SIR in a CDMA system. Pages 929–933 of: *Proceedings of the 47th IEEE VTC'97*, vol. 2.

Standard-95 TIA/EIA Interim. 1993. *Mobile station–base station compatibility standard for dual-mode wideband spread spectrum cellular systems*. Tech. rept. Telecommunications Industry Association.

R. S. Varga. 1962. *Matrix Iterative Analysis*. Englewood Cliffs, NJ: Prentice-Hall.

A. J. Viterbi, A. M. Viterbi and E. Zehavi. 1993. Performance of power-controlled wideband terrestrial digital communications. *IEEE Trans. Commun.*, 41(4), 559–569.

S. Yagnaraman. 2010. *Outer Loop Power Control for WCDMA*. Stockholm: KTH Royal Institute of Technology.

R. Yates. 1995. A framework for uplink power control in cellular radio systems. *IEEE J. Sel. Areas Commun.*, 13, 1341–1348.

J. Zander. 1992. Performance of optimum transmitter power control in cellular radio systems. *IEEE Trans. Veh. Tech.*, 41, 57–62.

7 Interference management

Radio spectrum is a finite resource, and therefore has to be shared by multiple users. Such *sharing* or *reuse* of radio spectrum is a dominant feature of present-day wireless systems where a tremendous number of devices have wireless connectivity. If you use mobile Internet, particularly in a city, it is highly likely that the frequency channel assigned to you is also used by someone very near you. This sharing of radio spectrum inevitably causes interference between users due to the broadcast nature of the wireless medium. Managing the interference between the wireless links is one of the most important problems in wireless networks.

We have already discussed some basics of interference management in Chapter 5. Let's briefly recall the cellular principle here. In a cellular system, the whole service area is divided into cells, and each cell is covered by a base station or an access point. The mobile terminals in the same cell are exempt from the interference problem because they are given orthogonal resources, e.g., frequency channel, time, spreading code, or waveform. However, interference still takes place between the neighboring cells. Thus, what matters in the cellular systems is *inter-cell interference*. Frequency planning is applied to the cells so that the proper *reuse distance* can be maintained between the cells of the same frequency channel assignment.

The static frequency reuse presented in Chapter 5 is of course not the only method of managing interference in practical wireless systems. In fact, immense research work has been performed and is still ongoing in the field of inter-cell interference management. In this chapter, we will go deeper into this subject. First, the categories and elements of interference management techniques will be introduced in Section 7.1. Then, in the subsequent sections, the three categories of interference management techniques, namely interference avoidance, interference randomization and interference cancellation, will be further discussed with examples. Finally, interference management for small cells and heterogeneous networks is presented in Section 7.5. Note that interference management is a vast research area. This chapter only scratches the surface of the field.

7.1 Classification of interference management techniques

We will give a brief overview of interference management before looking into specific techniques. In Section 7.1.1, a rough classification of the interference management techniques will be provided. In Section 7.1.2, the key elements of interference management will be presented. The advantages and disadvantages of the elaborated management schemes will also be discussed in this section.

7.1.1 Interference management categories

Existing interference management techniques can be roughly classified into the following three categories.

- **Interference avoidance**: this relies on the separation of users in space. Exponential decay of radio propagation enables us to utilize the same radio resource with a certain geographic separation, i.e., reuse distance. In order to make the separation in space, different times or frequencies can be allocated to the users in the vicinity of each other (time division or frequency division). Power control can also be adopted to adjust the required reuse distance. Simple forms of interference avoidance, e.g., static FDMA and TDMA, have been employed by most legacy cellular systems such as NMT and GSM. However, such rudimentary resource divisions usually lead to poor system performance because they cannot cope with traffic fluctuation. Efficient interference avoidance is still one of the main building blocks of the modern wireless networks.
- **Interference randomization**: this spreads the interference power over time or frequency in order to prevent the users from being hit by strong interference at a narrow frequency channel or at an instantaneous time. Examples of interference randomization include frequency hopping in Bluetooth and direct sequence spread spectrum in CDMA systems. With interference randomization, it is possible to build a system which is robust to a sudden fluctuation of interference or desired signal level. It is also useful when centralized resource allocation is difficult. However, the randomization becomes less effective as the resource utilization gets higher. This is because the interference cannot be spread sufficiently when there are not enough empty resources.
- **Interference cancellation**: this employs transmitter and/or receiver processing technologies to make the interference less harmful. The principle of interference cancellation can be described as follows: once interference is perfectly known, it is not harmful any more since it can be extracted from the received signal as if it did not exist. To do that, exact knowledge about the interference signal, e.g., instantaneous channel status information (CSI) and codeword, is required.* Although interference cancellation has great potential to significantly improve system performance, it suffers from the burden of channel estimation and reporting, backhaul communication for fast and reliable information exchange between access points, and the processing power of nodes. It is expected that advanced interference cancellation schemes will

* The term “channel” is used with two different meanings in the literature. First, it may refer to the unit of radio resource (e.g., frequency channel, channel allocation). Second, it may mean the condition of the wireless link due to propagation and fading (e.g., channel gain, channel fluctuation). The meaning should be distinguished by context.

be common features of practical systems in the near future with the help of powerful processors and optical fiber backhaul.

These categories will be discussed in more detail in the subsequent sections. It should be emphasized that the aforementioned interference management techniques are not mutually exclusive. Rather, you can combine the techniques in different categories to design more efficient interference management schemes. For example, you can group some adjacent cells as a bunch and employ interference cancellation techniques inside the bunch, while interference avoidance or randomization can be applied to mitigate the interference from distant cells.

7.1.2 Key elements of interference management

Numerous interference management schemes have been proposed so far, and even more schemes will probably be developed as the environments of network deployment and requirements for system performance evolve continuously. Some are simple, while others are more complicated. What kind of interference management scheme would be most beneficial for a wireless system? Is a complicated algorithm always better than a simpler one? These are not easy questions.

In general, complicated interference management algorithms utilizing more information about the interference characteristics are expected to provide better system performance. However, this is not always the case for the following reasons. First, the access points and terminals have to spend more time on measuring channel conditions since the instantaneous channel status varies quickly. Measured information may need to be reported to the transmitter side over the air, which also consumes radio resource. Those time and radio resources spent on using an algorithm may not be compensated if the interference level is not so high. Second, sophisticated schemes are usually vulnerable to estimation and reporting errors. It is not uncommon that an advanced scheme leads to worse performance in the presence of those errors than simpler schemes that are more robust to the errors.

Generally, you would have a better chance of performance improvement by employing a more complicated interference management scheme. However, there is another dimension to consider; such a performance improvement comes at the expense of increasing system complexity. Furthermore, it may incur additional investment for system operators. For example, inter-cell interference cancellation may demand the installation of inter-cell fiber optic connections and a powerful processor to perform joint signal processing. In this case, the cost benefit of the complicated scheme could be questionable. Therefore, the optimal algorithm giving the best performance is not always the best option for the system operators.

In order to choose or design an interference management scheme that suits the situation, you must figure out the key elements constituting the management strategies. Here we provide three key elements inspired by [R. Beck and H. Panzer, 1989].

- **Adaptivity to traffic**: When a system shows a long-term variation in average traffic load, it can be exploited by simple algorithms. For example, daily traffic patterns of office and residential areas are quite distinct and predictable. Large-scale resource lending and borrowing could be utilized. More sophisticated schemes can be adopted to handle instantaneous fluctuations in the traffic. Internet data traffic is usually elastic and bursty, and this dramatically increases the traffic fluctuation. This leads to a severe asymmetry of traffic loads between neighboring cells for short periods. Dynamic resource allocation could be employed if real-time traffic information is exchanged between the cells.

- **Adaptivity to channel conditions**: Radio propagation is mainly affected by terrain and the building landscape. Thus, average channel gain at a specific location does not change frequently. This enables us to design static or semi-static interference coordination schemes. On the other hand, instant channel status information (CSI) goes through a rapid variation. Utilizing the instantaneous CSI can potentially bring a huge performance improvement, but it comes with the burden of measurement, reporting and processing.
- **Resource reusability**: There exist a number of different algorithms for a certain level of adaptivity to traffic and channel conditions because there are several parameters that constitute an interference management scheme such as power, time, frequency, modulation and coding, and so on. A higher level of resource reusability means better system performance with a limited resource, which normally accompanies complex signal processing.

The elements of interference management are depicted in Figure 7.1. An interference management scheme of better quality will be needed when the traffic and channel variations are more dynamic and performance requirements are more stringent. In many cases, the best option satisfying both the performance requirements and the cost constraints could be found inside the cube in the figure.

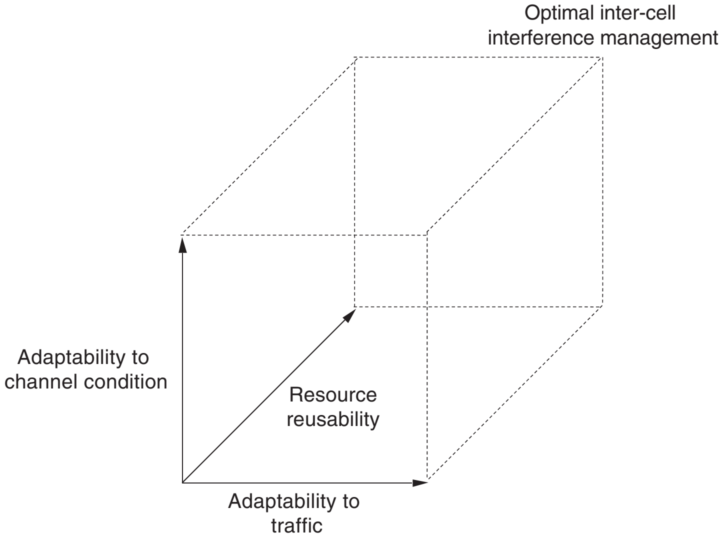

Optimal inter-cell interference management
Adaptability to channel condition
Resource reusability
Adaptability to traffic

Figure 7.1 Key elements of inter-cell interference management schemes.

7.2 Interference avoidance

Among the categories of interference management techniques discussed in Section 7.1.1, interference avoidance is the most commonly accepted in current systems. As mentioned earlier in this chapter, the basic way to handle inter-cell interference has been allocating different frequency channels to adjacent cells, i.e., frequency reuse, to maintain the reuse distance to meet the required level of SINR.

Although frequency reuse effectively lowers inter-cell interference, it also reduces the system capacity seriously because each cell can use only $1/K$ of the available spectrum when the system employs a cluster size (or frequency reuse factor) of K. Therefore, plenty of multi-cell resource allocation schemes have been proposed to achieve more efficient interference avoidance than plain frequency reuse. These schemes attempt to secure better adaptivity to channel conditions and/or to traffic. We will have a brief look at two distinct schemes: *reuse partitioning* and *multi-cell scheduling*.

7.2.1 Reuse partitioning

The idea of reuse partitioning emerged in the early days of cellular systems design as a remedy for the drawback of frequency reuse. Reuse partitioning is a static resource allocation scheme that exploits long-term interference characteristics of the cellular system.

Think of a mobile terminal moving from a cell center to a border. As it approaches the cell boundary, it receives a less strong signal from the serving access point, while the interference from the other cells gets stronger. Thus, its SINR drops sharply as it goes away from the serving access point. The frequency reuse factor should be determined to support the users at the cell border. This means that the terminals in the interior of the cell can tolerate a lower reuse distance. Instead of using a single frequency reuse factor throughout the system, several overlaid cluster sizes with different reuse distances can be applied to a cell. Terminals close to the cell center (high signal level) choose resources from a set for lower reuse factor, whereas those at the cell perimeter (low signal level) choose resources from a set with larger reuse distance.

Let us assume a cellular system with a regular hexagonal cell plan where all the access points use the same transmit power. The terminals use the same type of service (such as voice telephony) that requires the same amount of resource, namely a channel. A total of L different frequency reuse factors are used with cluster sizes of $K_1, K_2, \ldots, K_L$. Without loss of generality we number the reuse factors such that

$$K_1 < K_2 < \cdots < K_L. \tag{7.1}$$

A cluster with the reuse factor of K_i will have a reuse distance D_i given by

$$D_i = D_o\sqrt{3K_i}, \tag{7.2}$$

where D_o denotes the cell radius.

In order to study the basics of channel allocation in reuse partitioning, we neglect fading momentarily and only consider distance-dependent propagation loss with a path loss exponent of 4. Also, we assume that the system is interference limited. Thus, ignore background noise in the SINR calculation. Using these assumptions, the SIR may be computed for a terminal for distance d from the base station with full traffic load:

$$\Gamma(d) \approx \frac{{D_i}^4}{7.4d^4} = \frac{9{D_o}^4}{7.4d^4}{K_i}^2. \tag{7.3}$$

Note that this approximation holds when the reuse distance is not too small, i.e., when K is not small. However, we will employ it for small reuse factors as well for simplicity.

Assume that the terminals in the system require a certain SIR threshold ($\Gamma(d) \geq \gamma_0$). From this condition, we obtain d_i, the maximum distance within which the reuse factor K_i can be used:

$$d \leq D_o \left(\frac{9}{7.4\gamma_0} \right)^{1/4} \sqrt{K_i} = d_i. \tag{7.4}$$

Thus, a frequency reuse factor K_i can be used by all terminals up to a distance d_i from the cell center (the access point). A terminal in zone i may use a reuse factor with the cluster size K_i or larger. Figure 7.2 illustrates the overlay of the reuse clusters in a cell.

Having established K_i, satisfying the SIR requirement in various parts of the cell, we now turn to the actual resource allocation. This is done by simply determining the number of channels that is to be allocated to each zone. For this purpose, define the channel allocation vector as

$$c = (c_1, c_2, \ldots, c_L), \tag{7.5}$$

where the ith component c_i is the number of channels assigned to the terminals in cluster i. The composition of the channel allocation vector is constrained by the total number

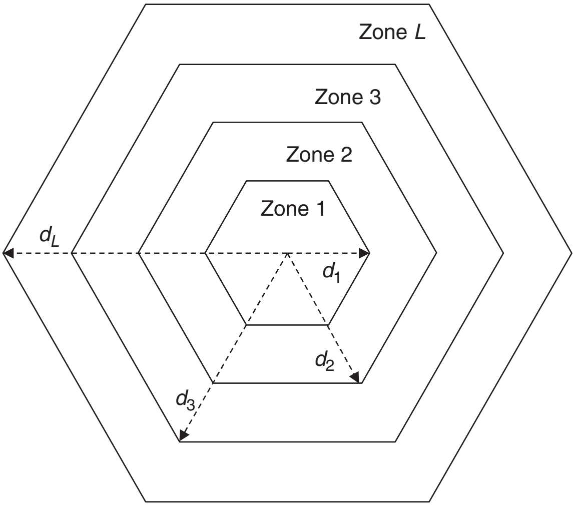


Figure 7.2 Clusters in reuse partitioning.

of available channels C with the following relationship:

$$\sum_{i=1}^{L} c_i K_i \leq C. \tag{7.6}$$

Considering the frequency reuse factors, the total number of channels available in a cell is given by

$$c_0 = \sum_{i=1}^{L} c_i. \tag{7.7}$$

Note that the terminals in zone i can utilize a reuse factor of K_i or larger, but cannot have channels of lower reuse factor. Therefore, only the terminals in the center of the cell will have access to all these c_0 channels whereas terminals on the perimeter are limited to using the c_L channels of the outermost zone.

There could exist several c vectors that satisfy the condition in (7.6). The optimal choice of c (for lowest assignment failure probability) depends on the distribution of the traffic inside the cell, i.e., allocating more channels where the traffic demands are greater. If the terminals are uniformly distributed, this is the same as choosing the c_i proportional to the size of the zone. Note that the size of zone i is proportional to ${d_i}^2 - {d_{i-1}}^2$. Also, from (7.4), we know that K_i is proportional to ${d_i}^2$. By defining $K_0 = 0$, we have

$$c_i = \frac{C(K_i - K_{i-1})}{\sum_{j=1}^{L}(K_j - K_{j-1})K_j}. \tag{7.8}$$

The reuse partitioning effectively splits the existing cell into smaller areas. The total number of channels available to the terminals increases because of the reduced reuse distances in inner zones. However, each zone has only a limited number of channels, which may lead to assignment failures at some zones even though the overall traffic load is not so high, i.e., trunking loss may occur. See the following example.

Example 7.1: Consider a cellular network with $C = 100$ channels available in the entire system. When traditional frequency reuse is applied, a cluster size of 9 is needed at the cell border to meet the SIR requirement, i.e., $K = 9$. The system employs reuse partitioning of five zones corresponding to $K = (1,3,4,7,9)$. With uniform distribution of mobile terminals, the optimal channel allocation vector of $c = (2,4,2,6,4)$ is obtained. This offers 18 channels per cell, which is considerably higher than the 11 channels that the conventional scheme provides. However, traffic variation in each zone may result in a higher assignment failure rate for low traffic loads. Calculate the assignment failure rates of the conventional channel assignment and reuse partitioning.

Solution: Recall that the assignment failure rate ν is defined as the number of assignment failures (Z) divided by the number of requesting terminals (M). For the case of reuse partitioning, this can be expressed as

$$\nu = \frac{\sum_i E[Z_i]}{E[M]} = \frac{\sum_i E[\max(0, M_i - c_i)]}{\omega A_c}, \tag{7.9}$$

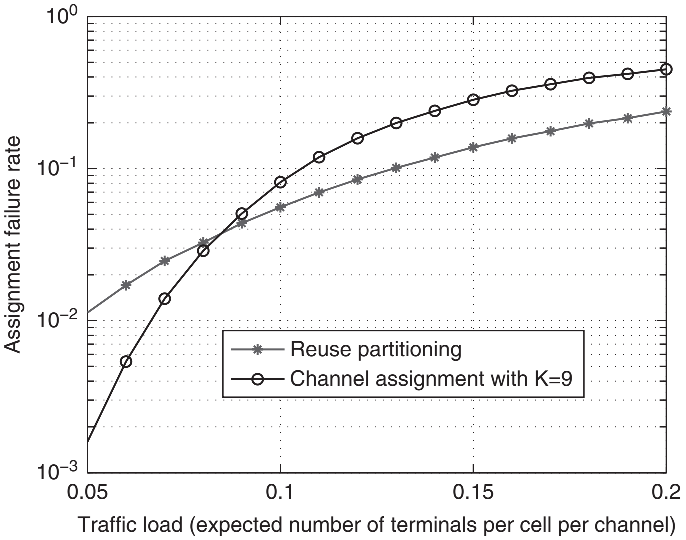


Figure 7.3 Assignment failure rate as a function of traffic load.

where Z_i is the number of assignment failures at zone i, M_i is the number of terminals at zone i, M is the total number of terminals in the cell ($M = \sum_i M_i$), and ωA_c is the expected value of M under the assumption that it is Poisson distributed.

We define $a = (a_1, a_2, a_3, a_4, a_5)$ as the vector denoting the portion of the terminals in each zone such that $\sum a_i = 1$. Since the terminals are uniformly distributed over the cell area, a_i is proportional to the size of zone i. This gives us $a = (1/9, 2/9, 1/9, 3/9, 2/9)$. Using the fact that M_i is Poisson distributed with the parameter $a_i \omega A_c$, the assignment failure rate ν can be obtained as

$$\nu = \frac{1}{\omega A_c} \sum_{i=1}^{5} \sum_{k=c_i}^{\infty} (k - c_i) \frac{(a_i \omega A_c)^k}{k!} e^{-(a_i \omega A_c)}. \tag{7.10}$$

The performance of the reuse partitioning scheme is illustrated in Figure 7.3 where the assignment failure rate is shown as a function of the traffic load ($\varpi = \omega A_c / C$). As expected, the assignment failure of reuse partitioning is higher than conventional frequency reuse with $K = 9$ when the traffic load is relatively low (when the average demand is below 11 channels). This example demonstrates that it is necessary to combine reuse partitioning with a traffic-adaptive channel allocation scheme to get good performance. An example of the adaptive scheme can be found in [J. Zander and M. Frodigh, 1992].

7.2.2 Multi-cell scheduling

The reuse partitioning in the above section is clearly more elaborate than the plain frequency reuse scheme. However, they both fall into the category of static resource allocation, which means they are not adaptive to traffic variation at a multi-cell level.

Once the channel allocation to each cell is determined in the system, no further interaction is made between the cells. The access points rely only on their own resources which have already been allocated to them, and do not care about what other cells are doing. This may perform well if the traffic loads are evenly distributed over the service area. In reality, however, we often observe uneven and dynamic variation of the traffic loads. For example, office districts and residential areas exhibit completely different daily traffic profiles. Furthermore, neighboring cells whose long-term traffic variations look alike often experience a big disparity in instantaneous traffic demands. Under the static schemes, it is not surprising to undergo a local deprivation of radio resource while neighboring cells have very little traffic at the same moment.

When communication links between the cells are poor and unreliable, it is necessary to implement a static inter-cell interference management scheme which minimizes inter-cell interactions. But what can be achieved if better inter-cell communication capability is available? If the cells can share their current traffic situations, it would be possible to adopt a dynamic resource allocation such that a cell with a high traffic load borrows some channels from cells with low traffic. The concept of *channel borrowing* attracted many researchers in the 1980s and 1990s. Interested readers are recommended to study [I. Katzela and M. Naghshineh, 1996] for more details about channel borrowing and other dynamic channel allocation schemes in the early days.

Let us further assume that the inter-cell connection is fast enough for the cells to exchange real-time scheduling information. Then why not make scheduling decisions together? This approach is called *multi-cell scheduling*. There are various ways of implementing the concept of multi-cell scheduling, one of which is depicted in Figure 7.4. Consider a group of neighboring cells. At each scheduling time, only one access point in the group is entitled to transmit. A scheduling manager (either one of the access points in the group or a higher entity supervising the access points) collects the real-time traffic information from the access points and determines which access point is eligible to transmit this time. The chosen access point can utilize all the radio resources, e.g., frequency channels, available for the group. Multi-cell scheduling can bring about a considerable improvement in system performance. First, inter-cell interference within the group is completely avoided. Second, there is no loss of efficiency due to the static

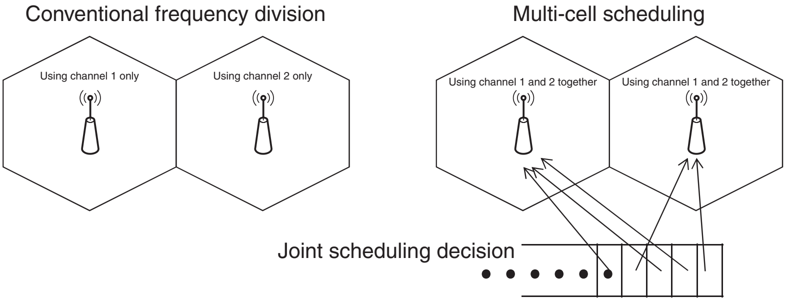

Conventional frequency division
Multi-cell scheduling
Using channel 1 only
Using channel 2 only
Using channel 1 and 2 together
Using channel 1 and 2 together
Joint scheduling decision

Figure 7.4 Multi-cell scheduling in comparison with conventional frequency reuse.

division of resources. The following example demonstrates the potential performance improvement.

Example 7.2: Consider a downlink of a cellular network consisting of three cells. It has a system bandwidth of 30 MHz. The access points receive data files from the Internet which are to be delivered to the mobile terminals. The arrival of the files follows a Poisson process with a rate of 10 files per second per cell, that is $\lambda_c = 10$. The file sizes follow an exponential distribution with an average of 1 megabit.

In an earlier configuration, the system used traditional frequency division with $K = 3$ to ensure a spectral efficiency of 2 bps/Hz throughout the system. This means that each cell exhibits a data rate of 20 Mbps (10 MHz × 2 bps/Hz). Since the file sizes are exponentially distributed, the delivery time of each file also follows an exponential distribution with an average of 0.05 second, i.e., $\mu_c = 20$. We can describe each cell in the network as an independent M/M/1 queueing system. By Little's law, the average time spent in the cellular network before files are delivered to the mobile terminals is

$$E(T) = \frac{1}{\mu_c - \lambda_c} = 100\, ms. \tag{7.11}$$

The system has adopted multi-cell scheduling with an upgrade of the backhaul link between the access points. The new configuration turns the whole system into one M/M/1 queue with a three-times faster service rate, i.e., $\mu_s = 3\mu_c$, and a system-wide file arrival rate λ_s. If the same delivery time is tolerated by the terminals, we can obtain $\lambda_s = 50$, which gives a per-cell arrival rate of 50/3. This means that the system can now deliver 67% more files to the terminals by implementing multi-cell scheduling.

Despite the rosy picture depicted in Example 7.2, it is difficult to realize the full potential of multi-cell scheduling in practical systems. A cellular network is typically comprised of thousands of access points. Applying multi-cell scheduling to a cluster of cells may solve the interference problem within the cluster, but the interference from/to the other cells surrounding the cluster still remains. The performance at the border of the cluster may even deteriorate because the terminals at the boundary may experience shorter frequency reuse distances to cells at the outside of the cluster. Therefore, inter-cluster interference should be mitigated in order to benefit from multi-cell scheduling.

7.3 Interference randomization

If an interference source spreads its signal energy over time or frequency, the impact of interference at a particular time instance or frequency channel will be reduced. One can say the interference is *randomized* (or *averaged*) if the spreading of the signal power is done with a random pattern. In this section, we will discuss interference randomization (also called interference averaging) based on *spread spectrum* techniques.

Examples of spread spectrum techniques are *direct sequence* and *frequency hopping*. In the direct sequence spread spectrum (DSSS) technique, the original bit stream is multiplied by a lengthy string of pseudorandom code which is called a pseudonoise (PN) sequence. The resulting signal looks like white noise to the receivers. However, the designated receiver which knows the exact PN sequence used by the transmitter is able to retrieve the original data by correlating the noise-like received signal with the PN sequence.

In the frequency hopping technique, the system bandwidth is divided into many narrowband frequency channels as in the case of FDMA. Also, time is split into slots similar to TDMA. A transmitter sends a narrowband signal on one of the channels in a time slot. In the subsequent time slots, the transmitter keeps using the same bandwidth, but selects a different channel at each slot. Thus, the transmission is made by hopping frequency channels. The selection of frequency channels is done according to a noise-like but predetermined PN sequence. When there are multiple communication pairs in the vicinity, their signals will randomly collide with each other as depicted in Figure 7.5. These random and occasional bit errors can effectively be restored by means of error correction coding.

The spread spectrum techniques have distinct advantages and disadvantages from the system design perspective. The benefits include robustness to fading, particularly to multipath fading, inherent encryption capability, and lower power spectral density. On the other hand, difficulty in synchronization is one of the drawbacks. CDMA technologies based on the DSSS technique have been used by commercial cellular systems for years, for taking advantage of DSSS in system design areas such as power control, voice activity factor and soft handover rather than for the interference randomization property alone [K. S. Gilhousen et al., 1991]. In this chapter, we will concentrate on the interference management aspect of the randomization.

Interference randomization is considered to be a practical solution for situations where the interference cannot be avoided effectively. First, it can be used in the presence of hostile interferers, e.g., jammers. Imagine a military situation where there is someone who wants to disturb your communication. If you stick to using a constant frequency

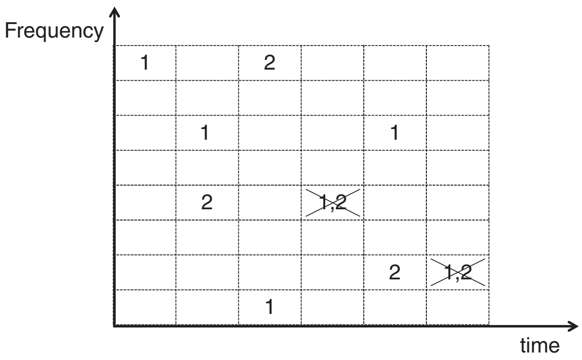


Figure 7.5 Time and frequency usage in a frequency hopping system with two active transmitters (Tx 1 and Tx 2); Slots with × mark denote collided transmissions.

channel, it would be easily detected, and your connection would be vulnerable to the jamming attack. When you cannot find a clean channel to avoid jamming, it would be better to randomly spread your signal because it would be difficult for the jammer to interrupt the entire bandwidth.

Second, it can be more efficient than simple resource division when traffic demand is bursty and relatively low. As demonstrated in Example 7.2, static resource division reduces the trunking efficiency of the system, which results in performance degradation. Exploiting knowledge about traffic demand can bring about more efficient resource utilization. However, it is not always possible to have real-time traffic information. Interference randomization can be an alternative way of maintaining the trunking efficiency provided that occasional bit errors can be taken care of by error correcting codes.

The third situation is when coordinated resource allocation is not available. It is not likely to happen in cellular networks, but personal communications in license-exempt spectra are a different story. Nowadays, virtually all personal and household devices have wireless connectivity. Think how many wireless communications are going on in a typical living room: wireless mouse and keyboard, wireless headset, game console, smartphone screen on TV, and so on. Most of these short-range communications are operated on the 2.4 GHz ISM band without a need for a spectrum license. They are not necessarily coordinated with the others. Some links may employ different air interface standards. In this situation, it is very difficult to avoid interference.

Example 7.3: In this example, we will consider a simplified depiction of multiple Bluetooth connections. Assume that there are M communication pairs in a small room. They are used for delay-sensitive applications such as video calls and music streaming. Thus, listen-before-talk MAC protocols such as CSMA are not an option for them. Instead, frequency hopping is applied to the pairs. The total bandwidth of 80 MHz is divided into 80 channels of 1 MHz each. Since the pairs are packed in a small space, a collision of slots always yields bit errors. However, the bit error rate of 10% is tolerated because of the error correction code. How many connections can coexist?

Solution: From our standpoint, a successful communication occurs when $M-1$ pairs happen to choose different channels than ours. Since there are 80 channels in total, the probability of a pair selecting a different channel is 79/80. Therefore, the bit error probability p_e is given by

$$p_e = 1 - \left(\frac{79}{80}\right)^{M-1} \leq 0.1. \tag{7.12}$$

Hence the maximum M satisfying the error probability constraint is nine.

If the communication pairs make random channel selections and keep using them constantly rather than doing frequency hopping, we obtain the same bit error probability. However, in this case, there will be some lucky pairs who enjoy error-free channels and others whose communications are completely impossible due to the consistent collisions. If one cannot risk a chance of connection failure, frequency hopping could be an attractive option.

7.4 Interference cancellation

It is well known that the propagation of radio waves is heavily affected by fading. This makes the received power of the signal fluctuate rapidly. The phase of the radio wave is also distorted due to fading. Accurate estimation of the wireless channel is thus of profound importance in wireless communications for proper demodulation of transmitted data. In cellular systems where you receive signals not only from your pair but also from several interferers, it is difficult to estimate the channels of all paths. Interfering signals are normally regarded as noise which, together with the background noise, hinders the demodulation of the desired signal. Therefore, we think interference is harmful.

Will the interference still be harmful if we manage to know its channel condition? Assume that accurate CSI of interfering signals is available to the receiver. Then these can be extracted from the received signal. It is also possible for the transmitters to manipulate the signal with the channel knowledge to make the receivers feel there is no interference. This idea sparked the concept of *interference cancellation*. Here we use the term interference cancellation to refer to a wide range of technologies that exploit channel knowledge and advanced signal processing rather than denoting a particular technique.

The basic idea of interference cancellation arose in the early 1970s [T. Cover, 1972]. Since then, there have been numerous studies suggesting that interference cancellation has great potential to improve the performance of wireless networks. However, it has not acquired much popularity in commercial systems. The major bottlenecks were the lack of processing power to handle complex signal processing and a slow backhaul connection inadequate for real-time cooperation between access points. Advances in microprocessors and optical networks have made interference cancellation techniques practically feasible in cellular systems. Therefore, the importance of these advanced techniques in wireless network design is expected to increase in the coming years.

Interference cancellation can be performed at the receiver side or transmitter side, or both. We will discuss examples of interference cancellation below: a receiver-oriented technique in Section 7.4.1 and processing at the transmitter side in Section 7.4.2.

7.4.1 Successive interference cancellation

Successive interference cancellation (SIC) is the most well-known technique at the receiver side. Consider a node receiving a mixed signal from $M+1$ transmitters. Only one of them sends the desired signal, and the others coming from M transmitters are interference. Let index 0 denote the desired signal. The interferers are ranked from 1 to M in order of decreasing received signal strength without loss of generality. The received signal y can be expressed as

$$y = \sum_{i=0}^{M} h_i x_i + n, \tag{7.13}$$

where x_i and h_i denote the transmitted symbol and channel gain of the ith transmitter, respectively; n is the background noise.

Assume that the receiver has accurate information about the channel of the strongest interferer, h_1. Then it can demodulate x_1, whose estimate is denoted by $\hat{x_1}$. With this, it can reconstruct the received signal by extracting $\hat{x_1}$ from the original signal:

$$y_1 = \sum_{i=0}^{M} h_i x_i + n - h_1 \hat{x_1}. \tag{7.14}$$

The receiver can continue, demodulating the second strongest interference and extracting it. In general, the reconstructed signal after the jth extraction is

$$y_j = \sum_{i=0}^{M} h_i x_i + n - \sum_{i=1}^{j} h_i \hat{x}_i. \tag{7.15}$$

If the demodulation of interference signals has been correct, y_j can be simply rewritten as

$$y_j = \sum_{i=0}^{M} h_i x_i + n - \sum_{i=1}^{j} h_i \hat{x}_i = h_0 x_0 + \sum_{i=j+1}^{M} h_i x_i + n. \tag{7.16}$$

In the end, only the desired signal x_0 will remain after M times of successful interference extraction.

Ideally, it is possible to eliminate all the interfering signals with M steps of SIC. In practical cellular networks, one or two strongest interferers normally contribute to the lion's share of inter-cell interference. Cancelling them can significantly improve the system performance even though a part of the interference still remains. For the SIC to work properly, it is important that the receiver has accurate estimates of the interference channels that it wants to cancel. Once an error occurs in the demodulation of the interference signal, the impact of the error will propagate to the rest of the cancellation steps and finally the demodulation of the desired signal. The error propagation is the major hurdle to tackle in the implementation of SIC.

7.4.2 Transmit beamforming

The burden of interference cancellation can be moved to the transmitter side if we assume that accurate CSI is known to the transmitters and fast backhaul communication between the transmitters is established. An example of such a technique is *precoding* or *transmit beamforming*. Consider a downlink of a cellular network where M access points deliver data to N terminals ($M \geq N$). All the access points are utilizing the same frequency channel, i.e., $K = 1$. The $N \times 1$ vector of received symbols $\mathbf{y}$ is denoted by

$$\mathbf{y} = \mathbf{h}\,\mathbf{x} + \mathbf{n}, \tag{7.17}$$

where $\mathbf{h}$ denotes an $N \times M$ matrix whose element h_{ij} is the complex channel gain between terminal i and access point j, $\mathbf{x}$ is an $M \times 1$ vector denoting the transmitted symbols, and $\mathbf{n}$ denotes a white noise vector with covariance $E[\mathbf{n}\mathbf{n}^{\dagger}]\sigma^2 = \mathbf{I}$. Note that the SINR for terminal i in the conventional cellular system is given by

$$\Gamma_i = \frac{p_i |h_{ii}|^2}{\sum_{j \neq i} p_j |h_{ij}|^2 + \sigma^2}. \tag{7.18}$$

The principle of transmit beamforming is that the access points cooperatively adjust the transmit powers to mitigate the impact of $\mathbf{h}$. It can be achieved by multiplying a prefilter matrix $\mathbf{A} \in \mathrm{C}^{M \times N}$ with the vector of desired signals $\mathbf{d}$ to build the vector of transmitted signals $\mathbf{x}$ ($\mathbf{x} = \mathbf{A}\mathbf{d}$). For this, the access points need to have an accurate estimate of the $\mathbf{h}$ matrix, i.e., channel gains between all the access points and terminals have to be known. Also, the access points must have a very fast backhaul connection to jointly construct the prefilter matrix $\mathbf{A}$ for every symbol transmission.

The simplest way of transmit beamforming is *zero forcing*, where nulling the interference is the prime objective of the prefilter matrix construction. Make $\mathbf{A}$ the pseudo-inverse of $\mathbf{h}$. That is,

$$\mathbf{A} = \mathbf{h}^{\dagger}(\mathbf{h}\mathbf{h}^{\dagger})^{-1}, \tag{7.19}$$

where $\mathbf{h}^{\dagger}$ denotes the Hermitian transpose of $\mathbf{h}$. With zero forcing, the received signal vector $\mathbf{y}$ becomes

$$\mathbf{y} = \mathbf{h}\mathbf{h}^{\dagger}(\mathbf{h}\mathbf{h}^{\dagger})^{-1}\mathbf{d} + \mathbf{n} = \mathbf{d} + \mathbf{n}. \tag{7.20}$$

Therefore, each terminal feels it receives an interference-free message from its own access point, though the received signal is actually a composite signal as a result of the cooperation of all access points.

Note that transmit beamforming requires network-level coordination [M. K. Karakayali et al., 2006]. This means that we need a reliable and high-speed backhaul over which the timely information about the prefilter matrix is conveyed and accurate channel estimation and reporting from the terminals. As discussed in Section 7.1.2, network coordination schemes such as transmit beamforming involve more complexity and cost. Therefore, an important question is whether the performance improvement is significant enough to compensate for the cost. There is no definitive answer to the question, but it depends on the environment and performance requirement [D. H. Kang et al., 2012]. With the processing power of devices increasing and optical fiber backhaul being more common, advanced network coordination is becoming an essential part of practical wireless systems. In Chapter 10, we will discuss network coordination features of the LTE system.

7.5 Interference management for heterogeneous networks

With the proliferation of high-end user devices such as smartphones and tablets, we are experiencing unprecedented data traffic growth. A forecast says that Internet traffic will almost double every year during this decade [Cisco, 2014]. It is a formidable challenge to provide the required capacity for the users of wireless and mobile communications. What makes the problem even more challenging is that the capacity provisioning must be achieved with only marginally higher or perhaps lower cost.

Deploying cheap and small access points to where traffic demands are higher (e.g. indoor offices and urban hotspots) is generally believed to be an integral part of the solution [J. Zander and P. Mähönen, 2013]. This will result in a layered deployment of large cells for coverage and small cells for more capacity, which is widely termed heterogeneous networks (HetNets). Figure 7.6 illustrates a typical configuration of a HetNet where small cells† are overlaid with macro cells.

There can be an enormous number of small cells deployed in the future, and they could interfere with each other and also with the macro cells. Therefore, interference management for HetNets is of significant importance. In principle, HetNets are not fundamentally different from the traditional cellular networks in the sense that any interference management scheme will be based on the three key elements: avoidance, randomization and cancellation. However, managing interference for HetNets is more difficult for several reasons [J. G. Andrews et al., 2012]. Many small cells are expected to be user-deployed, i.e. access points are installed by end users at any random places they want, which makes the interference unpredictable and occasionally very strong. Device price should be very cheap, and thus small cells are likely to have limited

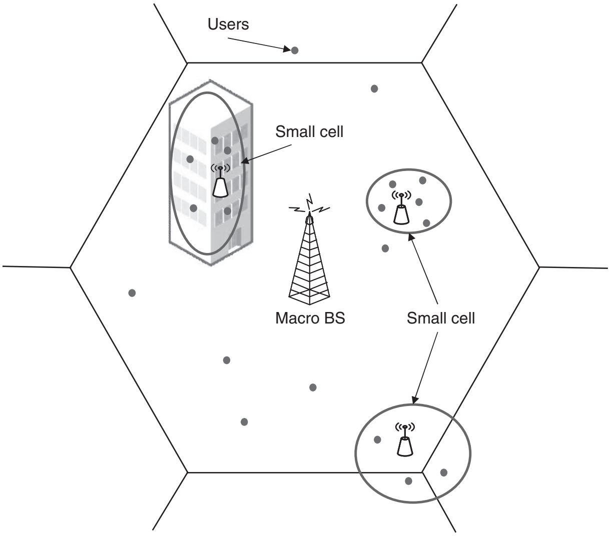


Figure 7.6 An illustration of a heterogeneous network configuration.

† These small cells are also referred to as femtocells in the literature.

interference management functionality. Access modes of small cells also affect the interference in HetNets. In the example below, we will discuss the impact of access modes in detail.

Example 7.4: In a HetNet, two types of access to small cells are considered:

- **Open Access**: any user in the network can have access to the small cell and benefit from it. See the left side of Figure 7.7. The open access mode is likely to be used in public places such as squares and shopping malls.
- **Closed Access**: access to the particular small cell is only allowed to users who are registered as a closed subscriber group (CSG). A non-CSG user may not be associated with the small cell even if he/she is very close to it. See the right side of Figure 7.7 where user 3 has to be connected to the macro base station although it is much closer to the small base station. The closed access mode can be common in private places such as residential buildings and offices.

Let us consider the downlink performance of users in Figure 7.7. Assume that all cells employ the same frequency channel with bandwidth $W = 10$ MHz. If there is more than one user in a cell, they share equally the available timeslots. Let $\mathbf{P} = \{P_{ij}\}$ be the received power from access point i to user j in dBm. In addition to the mutual interference, all users experience the same interference from outside of the macro cell, I_{ext}, with the power spectral density of –140 dBm/Hz. Background noise is neglected for simplicity. What is the achievable data rate for each user in each access mode? How is it different from a situation without a small cell?

$$\mathbf{P} = \begin{pmatrix} -65 & -60 & -70 & -80 \\ -90 & -50 & -70 & -80 \\ -75 & -70 & -40 & -60 \end{pmatrix}. \tag{7.21}$$

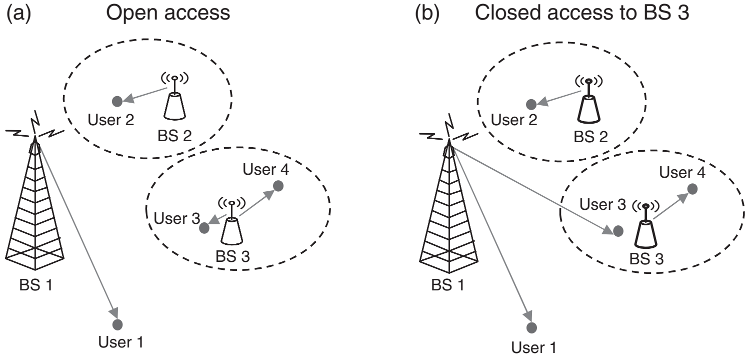


Figure 7.7 Example of open access and closed access.

Solution: Let s_j be the index of the serving access point for user j. Then the SINR of user j is given by

$$\Gamma_j = \frac{p_{s_j j}}{\sum_{i \neq s_j} p_{ij} + I_{ext}}, \tag{7.22}$$

where the value of I_{ext} is –70 dBm. Let n_i denote the number of served users at cell i. By applying Shannon's formula, the achievable data rate of user j is

$$r_j = \frac{W}{n_{s_j}} \log_2(1 + \Gamma_j). \tag{7.23}$$

When all users are served by the macro base station, the achievable data rates are $\mathbf{r} = \{5.14, 8.64, 2.50, 0.34\}$ in Mbps. In the open access, we can use $s_3 = 3$ and $n_3 = 2$. Then

$$\mathbf{r} = \{17.6, 32.2, 41.9, 16.1\}\ Mbps.$$

Significant capacity increase is achieved by installing the small cells. However, in the closed access, user 3 should be served by the macro cell and base station 3 becomes a source of severe interference, i.e. $s_3 = 1$ and $n_1 = 2$. It makes

$$\mathbf{r} = \{8.8, 32.2, 0.007, 32.2\}\ Mbps.$$

While some users still get reasonable data rates, the performance of user 3 is severely deteriorated. From the result, we observe that the closed access may lead to a serious performance degradation for non-CSG users who are close to small cells. Frequency division between macro and small cell layers can mitigate the problem to some extent. See the exercises for further details.

Exercises

7.1 A cellular communication system for voice telephony employs the reuse partitioning scheme with two zones as illustrated in Figure 7.8. Consider a cell with 100

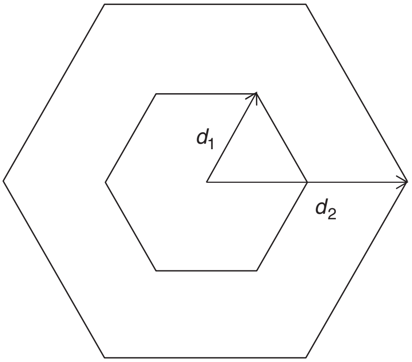


Figure 7.8 Reuse partitioning scheme with two zones.

channels available. Minimum SIR requirement is 16 dB. The cell radius is $d_2 = 200$ m. The path loss is distance dependent, proportional to the fourth power of the distance. For simplicity, fading effects are neglected.

a) Determine the frequency reuse factor required for the outer zone.
b) Obtain the maximum radius of the inner zone if we want to use a reuse factor of 3 in the inner zone.
c) Obtain the optimal channel allocation vector c under the assumption that the mobile terminals are uniformly distributed over the cell area.
d) Obtain the channel assignment failure rate as a function of the traffic intensity. (Assume that the traffic arrives with a Poisson distribution.)
e) Obtain the channel assignment failure rate for a conventional frequency reuse system and compare it with the performance of the reuse partitioning. What is the range of traffic load which makes the conventional system better than the reuse partitioning system?

7.2 Consider the downlink of a one-dimensional cellular system consisting of two identical access points with the coverage distance D_o as illustrated in Figure 7.9. Mobile terminal 1 is associated with access point 1, and mobile terminal 2 with access point 2. Terminal 1 and terminal 2 are assumed to be at the same distance x from their corresponding access points.

The two access points share 10 MHz of system bandwidth according to the following rule: if mobile terminals are closer to access points than a threshold distance d_{thr} (i.e. $x \le d_{thr}$), both access points transmit simultaneously on the full bandwidth; otherwise (i.e. $x > d_{thr}$), access point 1 and access point 2 divide the bandwidth into two.

In the following, employ the assumptions below for simplicity:

- The path loss exponent is 4. Fading effects are ignored.
- The noise power is 3 dB lower than the power received from one access point at the cell border (midpoint).
- The data rate according to Shannon's formula can be achieved in the system.

Then answer the following questions.

a) Express the SINR of mobile terminal 1 as a function of x.

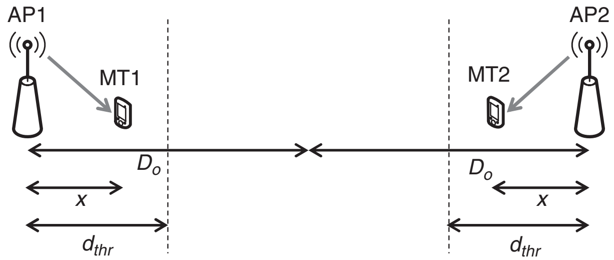


Figure 7.9 One-dimensional cellular system with two access points.

b) Assume that the mobile terminals are at the middle of their corresponding cells ($x = 0.5D_o$). Should d_{thr} be greater than $0.5D_o$ or not in order to obtain better performance in terms of the sum of data rates of the two mobile terminals?

c) When the mobile terminals are at the cell border ($x = D_o$), you will notice that resource division performs better ($d_{thr} < D_o$). What happens if the transmit power of the access points is reduced by half? Is the resource division still better at the cell border or not? (Notice that the noise power remains the same although the transmit power of the access points is changed.)

7.3 We discussed multi-cell scheduling in Example 7.2. In this problem, we will investigate a different multi-cell scheduling system. Consider a cellular network consisting of two cells. The cells use the same frequency band, and employ time division to avoid interference. Unlike Example 7.2, each cell performs its own scheduling due to insufficient backhaul. Instead, the following coordination is in place: Each cell area is divided into an inner zone and an outer zone. When both cells have picked users from the inner zone, they use one timeslot to transmit simultaneously. Otherwise, they allocate two timeslots for sending the packets, i.e., one slot for each. Assume that the cells always have data to transmit.

a) How much throughput increase can we expect compared to static time division? The inner zone radius is assumed to be $0.7D_o$.

b) What is the inner zone radius that gives a 30% throughput improvement over static time division?

7.4 Take Example 7.4 again and employ a simple frequency division such that half of the bandwidth is allocated to the macro cell and another half is used by the small cells. Assume that all the small cells fully utilize the allocated bandwidth. What is the performance under open access and closed access?

7.5 Consider a system consisting of four base stations as in Figure 7.10. The base stations are placed at an equal distance d from the point A, as illustrated in the diagram. The system employs a total bandwidth of 10 MHz. Your task is to evaluate the performance of different resource division schemes with or without multi-cell cooperation. There exist lots of combinations of resource division and cooperation schemes. Here we consider the three schemes described below:

- Scheme 1: frequency reuse of 1 without cooperation.
- Scheme 2: joint transmission of all four base stations with signal power combining.
- Scheme 3: cooperation of two-cell clusters, i.e. base stations 1 and 4 form cluster 1 and base stations 2 and 3 form cluster 2. Two clusters use the same frequency. Inside a cluster, joint transmission with signal power combining is adopted.

In the following, for simplicity we make the following assumptions:

- The path loss exponent is 2. Fading effects are ignored.
- All base stations have the same transmission power.

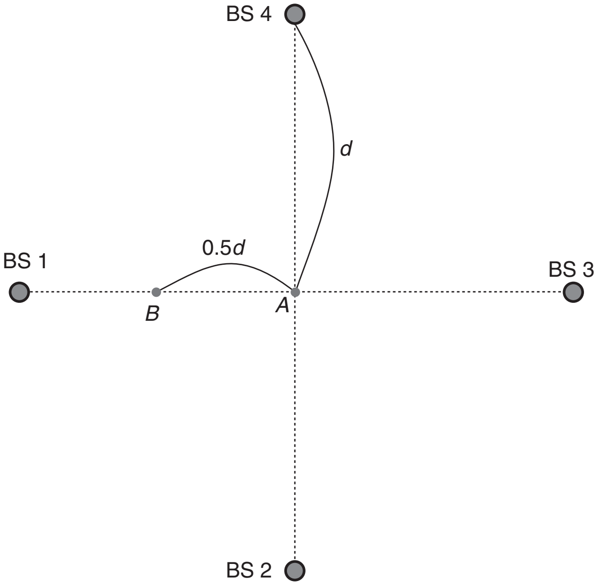


Figure 7.10 Network deployment.

- The noise power is 6 dB lower than the power received from one base station at the point A.

a) Consider a mobile at the point A. It is attached to base station 1. Identify the best and worst schemes from the viewpoint of achievable downlink capacity for the mobile. Assume that all cells are fully loaded, i.e. base stations 2, 3 and 4 also have their own users to serve. In each cluster, cooperating base stations employ an equal time division between them.

b) Consider a mobile at the point B who is attached to base station 1. Which one is better, scheme 1 or scheme 2?

c) Consider a mobile at the point B again. The difference is that not all base stations are fully loaded any more. Assume that base stations 3 and 4 momentarily have no user to serve. Which one is better, scheme 1 or scheme 3? Note that, for cooperation, base stations in a cluster do not need to share resources with non-loaded base stations.

References

J. G. Andrews, H. Claussen, M. Dohler, S. Rangan and M. C. Reed. 2012. Femtocells: past, present, and future. *IEEE Journal on Selected Areas in Communications*, 30(3), 497–508.

R. Beck and H. Panzer. 1989 (May). Strategies for handover and dynamic channel allocation in micro-cellular mobile radio systems. Pages 178–185 of: *IEEE 39th Vehicular Technology Conference (VTC'89)*, vol. 1.

Cisco. 2014 (June). *Cisco Visual Networking Index: Forecast and Methodology, 2013–2018*. White paper.

T. Cover 1972. Broadcast Channels. *IEEE Transactions on Information Theory*, 18(1), 2–14.

K. S. Gilhousen, I. M. Jacobs, R. Padovani, A. J. Viterbi, L. A. Weaver Jr. and C. E. Wheatley III. 1991. On the capacity of a cellular CDMA system. *IEEE Transactions on Vehicular Technology*, 40(2), 303–312.

D. H. Kang, K. W. Sung and J. Zander. 2012 (Dec. 3–7). Is multicell interference coordination worthwhile in indoor wireless broadband systems? In: *Proc. IEEE Global Communications Conference (GLOBECOM) 2012*.

M. K. Karakayali, G. J. Foschini and R. A. Valenzuela. 2006. Network coordination for spectrally efficient communications in cellular systems. *IEEE Wireless Communications*, 13(4), 56–61.

I. Katzela and M. Naghshineh. 1996. Channel assignment schemes for cellular mobile telecommunication systems: a comprehensive survey. *IEEE Personal Communications*, 3(3), 10–31.

J. Zander and M. Frodigh. 1992. Capacity allocation and channel assignment in cellular radio systems using reuse partitioning. *Electronics Letters*, 28(5), 438–440.

J. Zander and P. Mähönen. 2013. Riding the data tsunami in the cloud: myths and challenges in future wireless access. *IEEE Communications Magazine*, 51(3), 145–151.

8 Association and handover

If you want to make a phone call or surf the Internet with your mobile device, it must be connected to a base station as a first step. A cellular system consists of thousands of base stations. Thus, it is important to select the base station which can offer the best service quality. This is one of the fundamental radio resource management problems, namely *association*. When you are moving, the association has to be changed quite often because each base station has a limited coverage. This is termed *handover*.

In this chapter, we will study some basic problems involved with association and handover. We will start by looking at handover. In Section 8.1, terminology regarding handover types and procedure will be introduced. Then handover decision and resource allocation perspectives will be discussed in Section 8.2 and Section 8.3 respectively. Soft handover will be presented in Section 8.4. Finally, we will discuss association issues in heterogeneous networks (HetNets) in Section 8.5.

8.1 Anatomy of handover

8.1.1 Location management and handover

Nowadays, people expect seamless wireless connection while on the move almost everywhere in the world. The support of mobility is one of the main reasons behind the tremendous success of wireless communication systems. Two types of challenge need to be addressed in providing mobility support to users. The first one is keeping track of inactive terminals so that they can respond quickly to requests from the (fixed) network to establish communications with them. This is referred to as *location management* [V. W.-S. Wong and V. C. M. Leung, 2000]. It is of global scale, and sometimes involves multiple network operators because a terminal may leave the service area of one operator and enter that of another. This is called *roaming*.

The second challenge arises when an active terminal is moving. Since a base station only covers a limited area, the mobile terminal has a risk of leaving the area where its currently serving base station is capable of providing sufficient QoS. Therefore, mobility support for active terminals in cellular networks is achieved by timely and reliable transitions of serving base stations (*handover*). It should be performed in real time, and thus is a demanding task, particularly for delay-sensitive applications such as voice or video calls that require continuous service provision with very little loss of data. Handover is crucial for the QoS of not only the corresponding terminal but

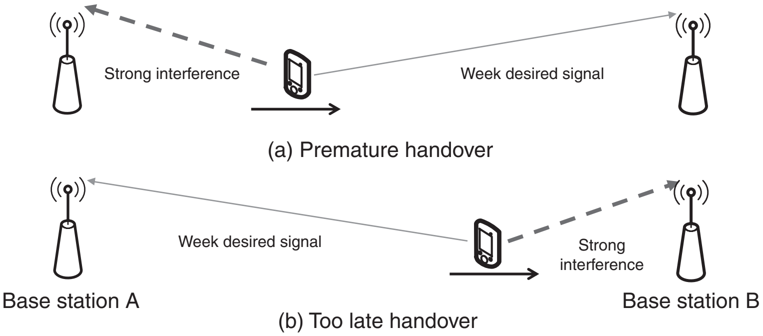


Figure 8.1 Problems of too early or too late handover.

also other terminals in neighboring cells. Figure 8.1 illustrates an uplink of a moving terminal. Premature or too late handover could cause a handover failure due to the low signal quality to the serving base station. At the same time, the neighboring cell may suffer from excessive interference since the terminal is too close to it.

8.1.2 Types of handover

Handover can be classified by the number of networks involved: *horizontal handover* and *vertical handover*.

- **Horizontal handover** is performed within a single network of homogeneous radio access technology (RAT). It is a central feature of any cellular network. In the remainder of this chapter, we will mainly deal with horizontal handover unless otherwise specified.
- **Vertical handover** is for a transition between different networks usually with heterogeneous RATs. Nowadays, a typical urban area is surrounded by a number of networks with various RATs which offer surplus coverage. Think of a shopping mall where you can have access to GSM, UMTS and LTE networks provided by several operators as well as WiFi hotspots. In this circumstance, vertical handover may occur even when a terminal remains stationary for reasons such as load balancing, higher data rate or cheaper price. This is also referred to as *network selection*. Mobile users want their devices to find and connect to the network offering the best speed, price or reliability, but they would not want to accept any disruption in services because of network selection. Smart and seamless vertical handover is an important issue for today's wireless networks. However, we will not cover vertical handover in this chapter. The challenges for vertical handover are mostly in network architecture and inter-networking protocols.

Horizontal handover can be divided into two types depending on the number of simultaneous connections, namely *hard handover* and *soft handover*.

- In **hard handover**, a terminal can be served by only one base station at any time. Thus, there could be an instant during which no base station supports the terminal while it is in transition between the base stations. In this regard, hard handover is also dubbed *break-before-make*. It has the advantage of simplicity. Most traditional cellular systems such as GSM have implemented hard handover, and even contemporary systems such as LTE adopt hard handover as the default handover method. However, it has a drawback of the so-called ping-pong phenomenon (or flip-flopping), i.e., very frequent switches between base stations, which could cause a large signaling burden on the system. It is necessary to prevent the ping-pong problem, but it may lead to an increase in handover failures. (See Section 8.2.2 for details.)

- In **soft handover**, the terminal can be served by two or more base stations simultaneously. The current and a few potential new base stations serve the terminal until it leaves the cell boundary region and is firmly attached to the next base station. Hence, it is called *make-before-break*. The difference between hard and soft handovers is similar to the difference between swimming relay events and track-and-field relay events. In swimming relays, the next swimmer starts just as the preceding one touches the wall, analogous to the switch from one base station to another in a hard handover. In track-and-field relays, the baton is passed from one runner to the next after the second runner starts running, and so for a short time they are both running together, analogous to a soft handover. The soft handover can remove the ping-pong phenomenon. Furthermore, macro diversity can be achieved by combining radio signals from the different base stations, which in turn results in better handover reliability and interference reduction. However, soft handover requires that the base stations use the same frequency band with a certain level of synchronization. It has been practically implemented in CDMA systems. (See Section 8.4 for details.)

8.1.3 Handover phases

Handover is about providing seamless services to active terminals while data transfer is in progress, i.e., timely selection and transition of the serving base station(s). Therefore, handover is a problem which directly relates to radio resource management. A handover can be divided into three phases:

- **Handover measurement and decision**: This refers to when and where to make a handover. In order to initiate the handover, a decision has to be made based on measurements of the current link quality or its predictions. Measurements are usually made at the terminal side. Then either the terminal or the network makes the handover decision, i.e. what the destination base station of the handover will be and when it will happen. Most cellular systems (e.g., GSM, UMTS and LTE) employ mobile-assisted (network-controlled) handover, namely MAHO, where the mobile terminal measures and reports the link quality levels to the serving base station, and then the base station or a higher entity makes a decision to initiate the handover. Mobile-controlled handover is another mode of operation where the terminal has the discretion, though this is not popular in contemporary systems. The handover decision is addressed in more detail in Section 8.2.

- **Handover resource management**: This phase concerns whether there are resources available in the target base station. When the target base station has been chosen and the decision for handover has been made in the previous phase, the question is whether there are enough radio resources in the destination base station. It is a crucial problem for delay-sensitive services such as voice and video calls because the calls have to be dropped if there is no available resource at the handover cell. This requires the base stations to reserve part of their radio resources for potential handover calls [I. Katzela and M. Naghshineh, 1996]. In traditional cellular systems where voice telephony accounted for the majority of services, handover resource management was one of the key challenges in overall network operation. For delay-tolerant data services, it has become less significant since the temporary deficiency of resources can be dealt with by allowing additional delay. Handover resource management will be discussed further in Section 8.3.
- **Handover execution**: This is about protocols and procedures for the reliable exchange of handover information. Assume that a target base station has been determined and it has enough resources to accommodate the terminal. Then a signaling procedure to implement the handover is needed. The terminal and base stations involved in the handover should be informed of the service termination at the current base station and new resource allocation at the new base station in a timely manner. Synchronization between the base stations may also be needed. The signaling procedure has to be fast and reliable in order not to lose the data and not to drop the connection while the handover is in progress.

This section has provided an overview of handover. In the remainder of this chapter, we will focus on the decision algorithms and performance evaluation of horizontal handover.

8.2 The handover decision problem

8.2.1 Performance metric

There could be several ways to evaluate the performance of a handover decision algorithm. Two of the most important metrics are described below:

- **Handover failure probability**: This is a measure of how frequently handover attempts fail. A failure is considered to occur if the signal quality is below a threshold for more than a certain time interval. Thus, the outage probability during the interval can be used as a good indicator of the handover failure in simple performance evaluations. The length of this interval is normally dependent on the service type offered to the terminal. Delay-sensitive services can tolerate only short interruptions. After a short moment, the service will be considered to be dropped. In delay-tolerant and bursty data services, a longer interruption can be withstood because packet transmissions can be delayed until the signal quality becomes acceptable.

- **Frequency of handover**: This measure illustrates how frequently handovers are executed. Consider Figure 8.2 where a terminal is moving straight from base station *A* towards base station *B*. If we do not consider fading in the radio propagation, only one handover will occur. In practice, it is highly likely that several more handovers are executed due to strong fluctuations in signal quality. However, some of the handovers may turn out to be unnecessary because the signal quality varies quickly. These excessive handovers normally impose an extra burden on the system signaling load, and thus can cause more handover failures. Note that it also implies that frequent handovers are not necessarily bad if the system has a signaling procedure to effectively handle them.

8.2.2 Tradeoff between handover frequency and failure

Let us focus on hard handover in this section. It is easy to speculate that there exists a tradeoff between the handover frequency and the failure probability. In order to draw a clear picture of the tradeoff, we assume a hypothetical system with accurate handover measurements, immediate and error-free handover execution, and enough radio resources in the destination cell. Then the handover performance is solely dependent on the handover decision with regard to the variation of the signal quality. Making fewer handovers increases the risk of handover failure. On the other hand, more handover decisions than necessary will increase the handover rate, which suggests an excessive signaling burden on the network in practical systems. We will take a closer look at this tradeoff with the following example [J. Zander and S.-L. Kim, 2001] where simple handover decision schemes are considered.

Example 8.1: Consider the mobile terminal in Figure 8.2. Assume it moves straight from base station *A* to base station *B* with a constant speed of 20 m/s. The distance between the base stations is 1000 m. The radio propagation loss is distance dependent with exponent α. It is also affected by a log-normal shadowing. In order to combat the

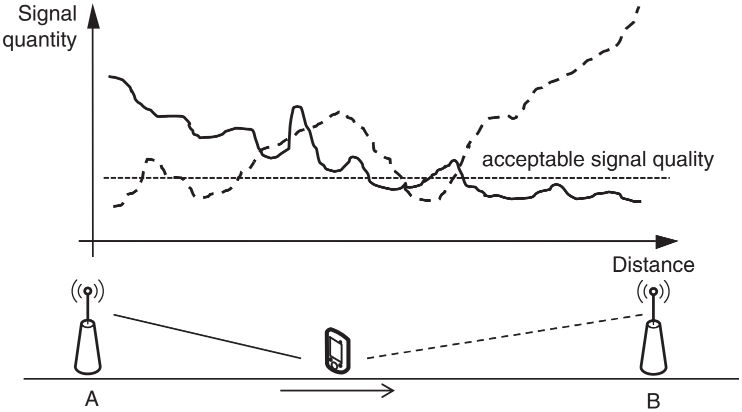


Figure 8.2 Fluctuation of signal quality for a moving terminal.

shadowing at the cell border (the mid-point between the base stations), the base stations transmit with a fading margin M_Δ so that the average received signal strength (RSS) is higher than the minimum required level. For the sake of simplicity, we consider the situation when there are only two base stations involved employing different frequency channels. Therefore, the impact of co-channel interference will be neglected. Under this assumption, the signal quality of the terminal can be described by the RSS from the two base stations. We assume that the base stations have unit transmission power without loss of generality. Then the RSS from base station i can be written as

$$S_i(d_i) = c\frac{L_i(d_i)}{d_i^\alpha}, \tag{8.1}$$

where d_i is the distance from the base station i, $L_i(d_i)$ is a log-normal random process, and c is the path loss constant. In the dB scale, we have

$$Y_i(d_i) = 10\log_{10}(S_i(d_i)) = C + X_i(d_i) - 10\alpha\log_{10}(d_i), \tag{8.2}$$

where $X_i(d_i)$ is a zero-mean Gaussian random process and C is the dB scale constant. The log-normal shadow fading will certainly be correlated such that $L_i(d_i)$ takes similar values for nearby points, i.e., $L_i(d_i) \approx L_i(d_i + \delta)$ for small distance δ. A simple but powerful model for describing this correlation can be found in [M. Gudmundson, 1991a,b].

Assume that the shadow fading has a standard deviation of 8 dB ($\sigma = 8$) and the correlation distance is 50 meters. Here correlation distance is defined as the distance where the correlation has dropped to 0.5. The terminal reports the signal level measurements every 0.5 seconds and the base stations can also make handover decisions at any moment of the reporting. Now determine

a) the outage probability,
b) the expected number of handovers.

Consider three simplified handover decision algorithms:

I. Instantaneous decision: the base station with higher RSS is chosen at any decision epoch.
II. Moving average: the RSS is averaged over the last ten measurements to determine the serving base station of this epoch.
III. Expected value: the handover is determined based on the expected value of RSS rather than the measurements.

Solution: The RSS from the two base stations is expressed in dB scale as follows:

$$\begin{aligned} Y_A(d_1) = &\quad C + X_A(d_1) - 10\alpha\log_{10}(d_1) \\ Y_B(d_1) = &\quad C + X_B(d_1) - 10\alpha\log_{10}(1000 - d_1) \end{aligned}, \tag{8.3}$$

where X_A and X_B are independent Gaussian random processes. Let $Y_{\min}$ be the minimum required signal level to maintain the connection. Since the base stations employ the fading margin M_Δ, we have $\mathbf{E}[Y_i(500)] = Y_{\min} + M_\Delta$.

As discussed before, the autocorrelation in the shadow fading can be expressed by using the model in [M. Gudmundson, 1991a,b]. In the model, it is assumed that the autocorrelation between k dB samples of the signal level taken at sample rate $1/T$ in a terminal moving at the speed v is given by

$$R_X(k) = \sigma^2 a^{|k|}, \tag{8.4}$$

$$a = {\epsilon_d}^{vT/d}, \tag{8.5}$$

where ϵ_d is the correlation of two shadow fading gains $X(r)$ and $X(r+d)$. In this example, we have

$$a = {\epsilon_d}^{vT/d} = 0.5^{20\times 0.5/50} \approx 0.87. \tag{8.6}$$

We can generate many realizations of the shadow fading processes using computer simulation. Figure 8.3 shows a typical example of the realization of the three algorithms. The instantaneous decision (algorithm I) means the mobile terminal is always connected to the base station from which it gets the best received power. Consequently, algorithm I will result in the best performance in terms of the outage probability. The moving average scheme (algorithm II) shows reasonable performance, but the averaging process causes a late handover decision as it tends to keep base station A as the serving base station even after passing the mid-point (500 m), which results in outage at the distance 500–750 m. The expected value scheme (algorithm III) allows only one handover at the mid-point. Although the decision makes sense from the viewpoint of average RSS, it leads to the lowest signal strength at many points due to the fluctuation of the received signal.

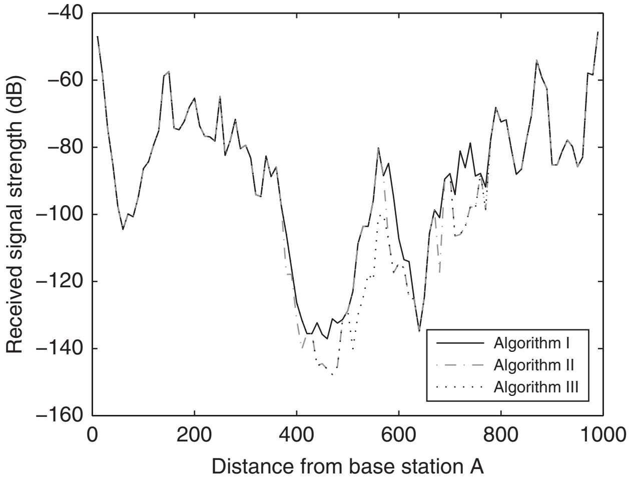


Figure 8.3 Typical realization of resulting signal levels for the three algorithms (I: solid line, II: dash-dot line, III: dashed line).

Table 8.1 Solution of Example 8.1.

Algorithm	N_{HO}	P_{out}
I	7.7	2.4 %
II	1.7	4.9 %
III	1.0	6.1 %

With a sufficient number of realizations, we can compute the outage probabilities and the handover counts. The results are shown in Table 8.1. Let's use $\alpha = 4$ and $M_{\Delta} = 5$dB. As expected, the instantaneous decision provides the best results in terms of the outage probability, whereas the expected value scheme performs the worst in this situation. The tradeoff between the handover frequency and failure is also clearly visible. The number of handovers, N_{HO}, is very large for algorithm I. For algorithm III, only one handover always occurs at the cell border, leading to the worst outage performance.

Example 8.1 clearly shows the importance of an effective smoothing procedure on the instantaneous measurements in order to strike a good balance between the number of handovers and the outage probability. A simple moving average scheme is used in this example for illustrative purposes. However, a number of better techniques utilizing knowledge about the fading environment have been proposed. If we can predict what the signal strength will look like, we will be able to make a proactive handover decision before the signal quality deteriorates too much.The drawback of such sophisticated prediction techniques is that they need information about radio propagation and fading parameters, which vary a lot in real situations and are difficult to know. Thus, the potential of the prediction techniques is difficult to realize.

For practical systems, a simple nonlinear technique from control theory, i.e. signal hysteresis, has commonly been used to reduce the flip-flopping of handovers. With hysteresis, the handover decision is made when the difference between two signals (from base station A and base station B) is above a certain threshold. Assume a handover from A to B is decided. This means $Y_B > Y_A + \zeta$, where ζ denotes the hysteresis threshold. Going back to A is only possible when $Y_B + \zeta < Y_A$. Therefore, the requirement for "going back" becomes higher.

8.2.3 Impact of handover criteria

In the previous discussion, the received signal level was used as the handover decision criterion. However, this is not always adequate because inter-cell interference may affect the signal quality significantly in most urban areas, as can be seen in the next example.

Example 8.2: The drawback of using RSS as the decision criterion is noticeable in dense-urban-like propagation conditions such as in the Manhattan model [O. Grimlund and B. Gudmundson, 1991; J. Zander and S.-L. Kim, 2001]. Figure 8.4 depicts an example. Here four base stations are located in a downtown area. Base station 2 is using the same frequency channel as base station 1, whereas the other two base stations utilize another channel. Further, assume that the signal levels are reasonably high all over the area.

We will qualitatively examine the handover in the presence of rapid signal level changes as a moving terminal turns around street corners. Assume a terminal on a street close to base station 1 in the lower middle of the map in Figure 8.4. It starts moving towards base station 2, keeping straight on path A, and taking a right turn towards base station 4 on path B. Two handover decision criteria, namely RSS and SIR, are considered as illustrated in Figure 8.5.

For path A, the signal level (A-RSS) drops gradually, whereas the SIR (A-SIR) decreases swiftly due to the inter-cell interference. Therefore, the RSS-based scheme cannot notice that a handover is urgent. For path B, after the mobile takes the turn, a dramatic drop in signal level (B-RSS) is observed in Figure 8.5 as the mobile is affected by the "building shadow" immediately after the turn. Thus, the RSS-based decision will be to make a handover as soon as possible. On the contrary, the SIR trace of the mobile (B-SIR) tells a different story. As the main interferer is base station 2, the desired signal and the interference decreases by about the same, which makes their ratio, SIR, almost unchanged. Handover will not be necessary because the SIR is still in a good range.

Example 8.2 demonstrates that the RSS should not be the sole handover decision criterion, particularly in dense urban areas where interference is the issue rather than

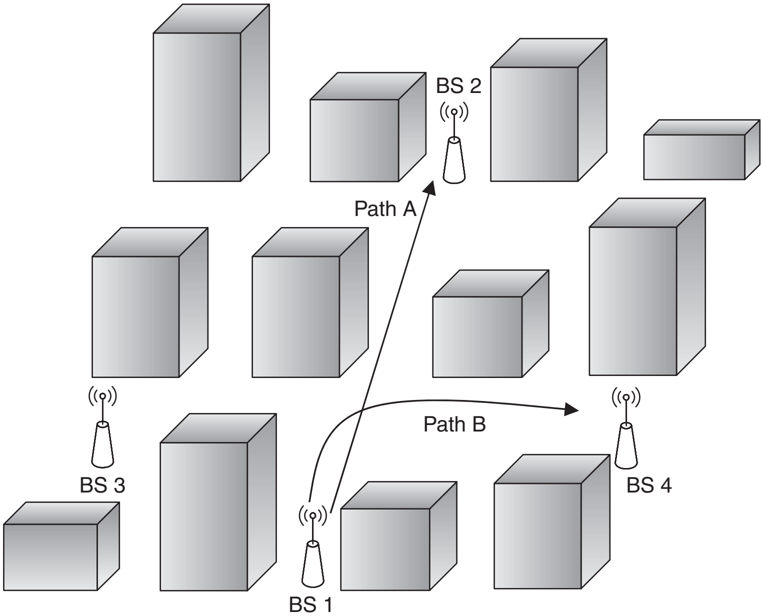


Figure 8.4 A Manhattan-like propagation environment with four base stations.

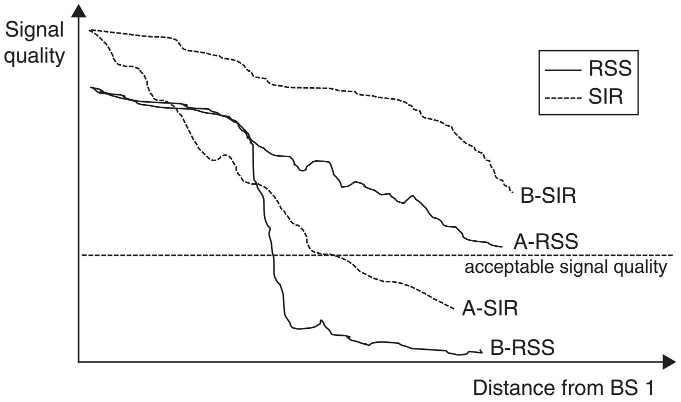


Figure 8.5 RSS vs SIR as handover decision criteria.

the coverage. In practical systems, both RSS and SIR are used depending on the interference situation.

8.2.4 An example handover decision algorithm

In this section, we have a brief look at a typical hard handover decision algorithm [S. Sesia et al., 2009]. The description is based on the handover in LTE systems, but the algorithm is not specific to LTE. Basic handover operations in most cellular systems are quite similar. In principle, MAHO is employed such that the terminal measures and periodically reports the signal quality, and the base station makes a decision on when and where to hand over. In LTE systems, the terminal measures cell-specific reference signal received power (RSRP). This measurement is performed to determine the ranking of candidate cells in terms of the signal strength. Additionally, the terminal may measure the received signal strength indicator (RSSI), which is the sum of total received power. Then reference signal received quality (RSRQ) can be calculated as the ratio between RSRP and RSSI. The RSRQ measurement is used when RSRP does not provide sufficient information to make a reliable handover decision, for example in the circumstances discussed in Section 8.2.3.

The LTE handover algorithm can be described by the following key concepts: *hysteresis* and *time-to-trigger*. Hysteresis refers to an offset of measured signal qualities between serving and target cells. A handover is initially considered if the measured RSRP or RSRQ of the target cell is greater than that of the serving cell with the offset of hysteresis value. Then actual handover is performed when this condition holds for a certain duration, i.e., time-to-trigger.

The hysteresis and time-to-trigger can effectively prevent the ping-pong problem. However, choosing the right parameters for the hysteresis and time-to-trigger is not a trivial task. Those parameters should be optimized according to the deployment of base stations, propagation environment, and traffic and mobility behaviors of the terminals. Figure 8.6 depicts the described handover decision algorithm.

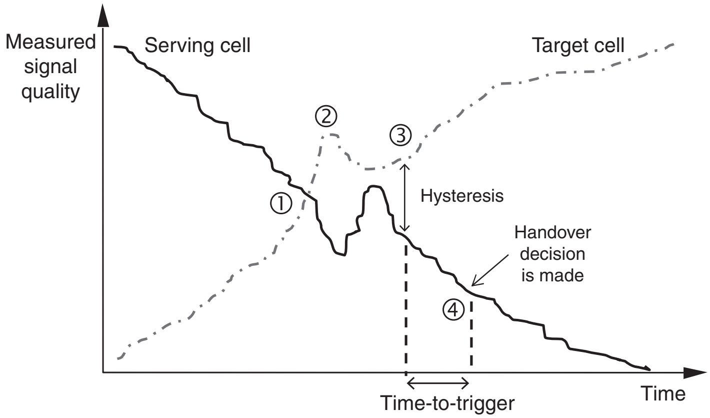


Figure 8.6 An illustration of the handover decision algorithm for LTE systems.

Example 8.3: In Figure 8.6, the handover decision is made at time point 4. What prevented the handover decision at the previous time points?

Solution: At point 1, the signal quality of the target cell became better than that of the serving cell, but the difference was still less than the hysteresis. At point 2, the hysteresis condition was met, but it did not sustain. The hysteresis was satisfied again before point 3, but time-to-trigger is only fulfilled at point 4, where the decision is eventually made.

8.3 Handover resource management

When a handover decision has been made, the base station of the target cell should be ready to accept the terminal. This means that the new serving cell must have enough radio resource to handle the traffic of the incoming terminal. If there is no resource available, the service will be delayed, and may even be terminated unwillingly in the worst situation.

When a cell is heavily loaded, it is not only handover terminals who suffer from the lack of resource. Mobiles trying to set up a new connection will also have difficulties in obtaining the required resources. Thus, prioritization between the handover users and the new users becomes an issue. It is generally considered that protecting ongoing sessions is more important than accepting new sessions to the system. People would be more annoyed when ongoing phone calls are suddenly dropped than when new call attempts fail. Therefore, it is necessary to reserve some resources for handover terminals (so called *guard channels*) even if it leads to increased rejections of new service requests.

It is obvious that delay-sensitive services are much more susceptible to the resource availability of the target cell than delay-tolerant ones are. In legacy cellular systems where voice telephony accounted for the majority of the services, handover resource management, combined with admission control of new calls, was one of the central problems in overall radio resource management. The significance of handover resource management has diminished over the years with the proliferation of Internet applications. For instance, a few seconds of disruption would not be a problem for watching a YouTube video as long as enough bits have been buffered already. However, delay-sensitive and resource-demanding services are also a part of mobile data traffic. Imagine you watch a live football match through your mobile device. You would not be happy if you get a video with seconds of delay and are susceptible to a spoiler from a friend.

We will study the cost of handover resource reservation in Example 8.4, where a delay-sensitive service is assumed. Then we will see how delay-tolerant services can ease the situation in Example 8.5.

Example 8.4: Consider a cellular system where each cell has N_{tot} channels. It is assumed that the terminals in the system employ a homogeneous delay-sensitive service, e.g., video calls. In a cell, new calls and handover calls arrive as independent Poisson processes with intensities λ_n and λ_h, respectively. The traffic demand in the cell is defined as $\lambda_t = \lambda_n + \lambda_h$. The calls have a lifetime (or sojourn time) in the cell, which is the time spent before they are terminated or leave the cell. Call lifetime is exponentially distributed with an average of $1/\mu$.

In order to prioritize handover calls, N_{rsv} channels are exclusively reserved for handover. Thus, a new call is not admitted (and only handover calls are accepted) when the number of currently used channels is equal to or larger than $N_{thr} = N_{tot} - N_{rsv}$. However, handover calls are dropped only when all N_{tot} channels are occupied.

Assume that 50% of total calls are accounted for by handover calls (i.e., $\lambda_n = \lambda_h$). A total of $N_{tot} = 12$ channels are assigned to the cell. Further, $\mu = 1$. Determine the new call blocking probability and handover call dropping probability as a function of traffic load (λ_t) and the number of reserved channels ($N_{rsv} = N_{tot} - N_{thr}$).

Solution: Let $N(\tau)$ denote the number of ongoing calls in the cell at time τ. Since the call arrivals follow independent Poisson processes and call durations are exponentially distributed, $N(\tau)$ can be described as a birth–death process as depicted in the state transition diagram in Figure 8.7. Our task is then to find out steady-state probabilities p_k such that

$$p_k = \Pr[N(\tau) = k],\ 0 \le k \le N_{tot}. \tag{8.7}$$

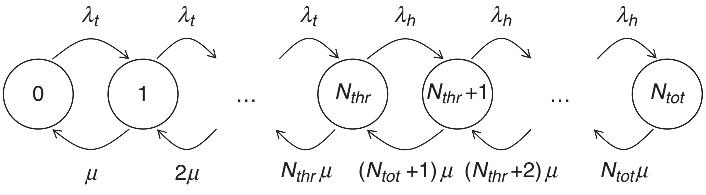


Figure 8.7 State transition diagram for Example 8.4.

Note that the new call blocking probability p_{block} and handover call dropping probability p_{drop} are given by

$$p_{block} = \sum_{k=N_{thr}}^{N_{tot}} p_k, \tag{8.8}$$

$$p_{drop} = p_{N_{tot}}. \tag{8.9}$$

From Figure 8.7, we can derive the following balance equations:

$$\lambda_t p_{k-1} = k\mu p_k, \quad 1 \leq k \leq N_{thr}, \tag{8.10}$$

$$\lambda_h p_{k-1} = k\mu p_k, \quad N_{thr}+1 \leq k \leq N_{tot}. \tag{8.11}$$

p_k can be derived by iteratively solving these equations. Then we get

$$p_k = \begin{cases} p_0 \dfrac{{\lambda_t}^k}{\mu^k k!} & k \leq N_{thr} \\ p_0 \dfrac{{\lambda_t}^{N_{thr}} (\lambda_h)^{k-N_{thr}}}{\mu^k k!} & N_{thr}+1 \leq k \leq N_{tot} \end{cases}. \tag{8.12}$$

We can obtain p_0 by using the fact that $\sum p_k = 1$. Define the following notations: $\rho_t = \frac{\lambda_t}{\mu}$ and $\rho_h = \frac{\lambda_h}{\mu}$. Then

$$p_k = \begin{cases} \dfrac{\dfrac{{\rho_t}^k}{k!}}{\sum_{j=0}^{N_{thr}} \dfrac{{\rho_t}^j}{j!} + \sum_{j=N_{thr}+1}^{N_{tot}} \dfrac{{\rho_t}^{N_{thr}} {\rho_h}^{j-N_{thr}}}{j!}} & k \leq N_{thr} \\ \dfrac{\dfrac{{\rho_t}^{N_{thr}} {\rho_h}^{k-N_{thr}}}{k!}}{\sum_{j=0}^{N_{thr}} \dfrac{{\rho_t}^j}{j!} + \sum_{j=N_{thr}+1}^{N_{tot}} \dfrac{{\rho_t}^{N_{thr}} {\rho_h}^{j-N_{thr}}}{j!}} & N_{thr}+1 \leq k \leq N_{tot} \end{cases}. \tag{8.13}$$

Figure 8.8 shows the numerical results. The tradeoff between the new call blocking and handover dropping is clearly observed in the figure. In order to achieve lower p_{drop}, more channels have to be reserved, which in turn leads to higher p_{block}.

Example 8.5: In this example, we will see if channel reservation is still necessary for delay-tolerant applications. Consider the cellular system of N_{tot} channels per cell again. Mobile terminals use a homogeneous application, but this time it is a delay-tolerant application such that a temporal service interruption is allowed during handover and new call attempts. We use the same parameters for call arrivals and lifetime: λ_n, λ_h and μ.

Unlike the previous example, we do not prioritize handover calls, i.e., no channel reservation. This means that p_{block} will be equal to p_{drop}. When there is no channel available at the moment of (either new or handover) call arrival, the terminal can be

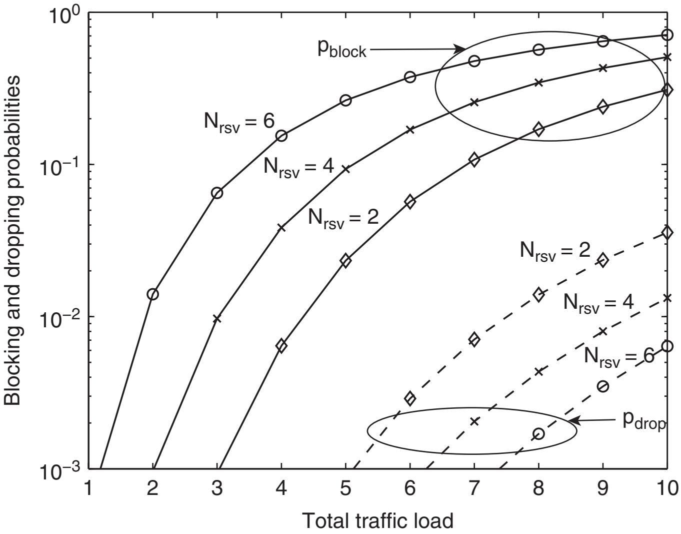


Figure 8.8 Blocking and dropping probabilities as functions of traffic load and reserved number of channels.

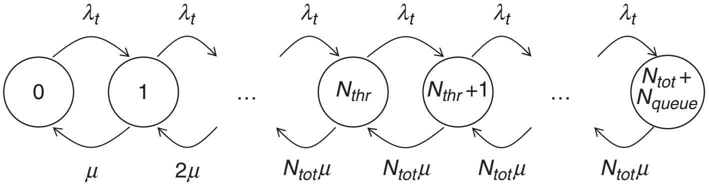


Figure 8.9 State transition diagram for Example 8.5.

put into a queue so that it can wait until it finds a vacancy. The queue size is denoted by N_{queue}. The larger the queue size is, the more delay-tolerant the users are. For the assumption of $\lambda_n = \lambda_h$ and $N_{tot} = 12$, determine p_{drop} as a function of the traffic load and the queue size.

Solution: We can follow the steps used in Example 8.4 with a slight modification. First, let us draw a state transition diagram for a birth–death process (see Figure 8.9). Note that there are N_{queue} more states after the state N_{tot}. Balance equations are derived as follows:

$$\lambda_t p_{k-1} = k\mu p_k, \quad 1 \le k \le N_{tot}, \tag{8.14}$$

$$\lambda_t p_{k-1} = N_{tot}\mu p_k, \quad N_{tot}+1 \le k \le N_{tot}+N_{queue}. \tag{8.15}$$

By iteratively solving these equations,

$$p_k = \begin{cases} p_0 \dfrac{{\lambda_t}^k}{\mu^k k!} & k \le N_{tot} \\ p_0 \dfrac{{\lambda_t}^k}{\mu^k N_{tot}^{k-N_{tot}} N_{tot}!} & N_{tot}+1 \le k \le N_{tot}+N_{queue} \end{cases}. \tag{8.16}$$

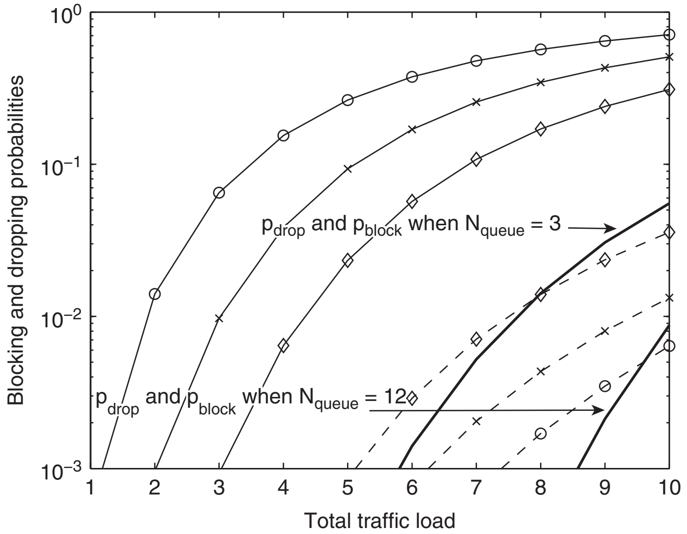


Figure 8.10 Blocking and dropping probabilities when a delay-tolerant application is considered.

Thus,

$$p_k = \begin{cases} \dfrac{\dfrac{{\rho_t}^k}{k!}}{\sum_{j=0}^{N_{tot}} \dfrac{{\rho_t}^j}{j!} + \sum_{j=N_{tot}+1}^{N_{tot}+N_{queue}} \dfrac{{\rho_t}^k}{{N_{tot}}^{k-N_{tot}} N_{tot}!}} & k \leq N_{tot} \\ \dfrac{\dfrac{{\rho_t}^k}{{N_{tot}}^{k-N_{tot}} N_{tot}!}}{\sum_{j=0}^{N_{tot}} \dfrac{{\rho_t}^j}{j!} + \sum_{j=N_{tot}+1}^{N_{tot}+N_{queue}} \dfrac{{\rho_t}^k}{{N_{tot}}^{k-N_{tot}} N_{tot}!}} & N_{tot}+1 \leq k \leq N_{tot}+N_{queue} \end{cases} . \quad (8.17)$$

Figure 8.10 shows the resulting dropping probability in comparison with Example 8.4. We can notice that introducing a small queue significantly lowers the risk of handover drop. Moreover, new calls do not need to be discriminated against any more in this example.

8.4 Soft handover

So far, we have studied the characteristics of hard handover. This section will deal with soft handover. As indicated earlier in the chapter, soft handover is a technique whereby a terminal in transition between one cell and its neighbor transmits and receives the same signal to/from both base stations simultaneously, as illustrated in Figure 8.11. Soft handover can occur between two or more base stations.

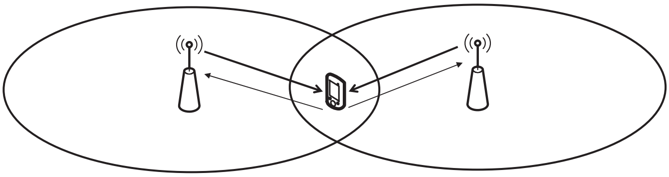


Figure 8.11 An illustration of soft handover.

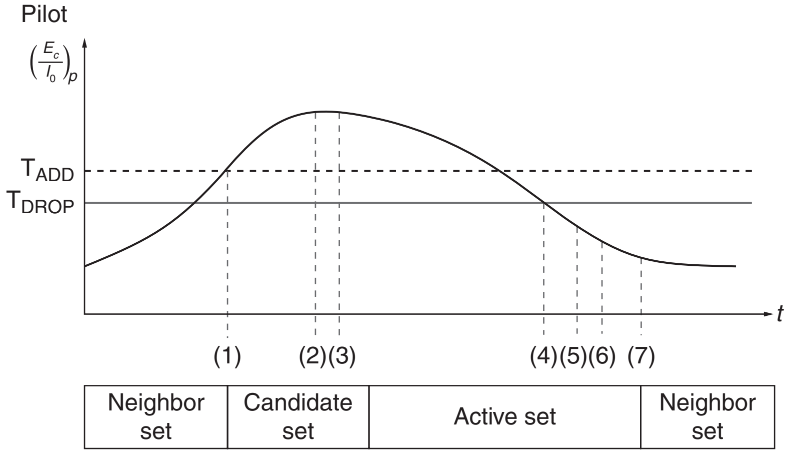


Figure 8.12 Soft handover in the IS-95 CDMA system.

8.4.1 Soft handover procedure in practical systems

Soft handover has been used by DS-CDMA systems where terminals can reduce the transmission power by employing soft handover and the total interference of the system can be decreased. In soft handover mode, a terminal receives power control commands from different base stations. It is possible that base stations give different commands, e.g. one sends power-down commands and the other sends power-up commands. This probably happens when the terminal is approaching a new base station. In this situation, the terminal will decrease its power since it is possible to communicate with a reduced transmission power. DS-CDMA systems employ soft handover on both the downlink and the uplink. On the downlink, the rake receiver at the mobile terminal can achieve maximum ratio combining (MRC) of the two incoming signals. On the uplink, the mobile switching center must resolve which base station is receiving the stronger signal and decide in its favor (selection diversity).

In the IS-95 system, the soft handover decision is based on the pilot strength measurement and maintenance. The set called *active* includes the base stations currently interacting with the terminal. Therefore, if the cardinality of the active set is greater than one, it means that the terminal is in soft handover. As illustrated in Figure 8.12, there are three thresholds that will determine the soft handover status: T_{ADD}, T_{DROP} and a timer interval T_{TDROP}. If the received E_c/I_0 (chip energy per interference spectral density) of a pilot channel of a base station in the candidate set exceeds T_{ADD}, the mobile terminal sends a pilot strength measurement message to the base station and transfers the pilot to the candidate set. The base station sends a handover direction message to the mobile terminal. The mobile terminal transfers the pilot to the active set and sends a handover completion message, that is a new connection is established. If the received E_c/I_0 of a pilot channel of a base station in the active set drops below T_{DROP}, the mobile terminal starts the handover timer. If the timer expires (T_{TDROP}) and the pilot E_c/I_0 is still below T_{DROP}, then the mobile terminal sends a pilot strength measurement message to the base station; and the base station sends a handover direction message back. The mobile terminal then moves the pilot from the active set to the neighbor set and sends a handover completion message to the base station; that is, the connection to the base station is removed.

In IS-95 soft handover, we can adjust the soft handover region by changing the values of T_{ADD} and T_{DROP}. Increasing T_{ADD} and T_{DROP} will decrease the soft handover region and decreasing them will have the opposite effect. The same phenomenon will happen if the received E_c/I_0 from both base stations changes instead. Since the received E_c/I_0 is a function of the interference experienced within the system, any load variation within the IS-95 system will affect the soft handover region. A light traffic load leads to a larger soft handover region and a heavy load situation will shrink the soft handover region. Therefore, the soft handover rate in IS-95 is load dependent and changes according to the traffic density.

To avoid this phenomenon, the WCDMA system uses a slightly different mechanism. Here relative handover thresholds between received pilots are employed, as illustrated in Figure 8.13. When an incoming base station's pilot channel E_c/I_0 is larger than the current base station by T_{ADD}, the soft handover starts with the new base station. Also, when the current base station's pilot channel E_c/I_0 is less than that of the incoming base station by T_{DROP}, the current base station is removed from the connection. As depicted in Figure 8.13, the soft handover region in WCDMA is independent of the traffic load since any interference change will push both pilots up or down.

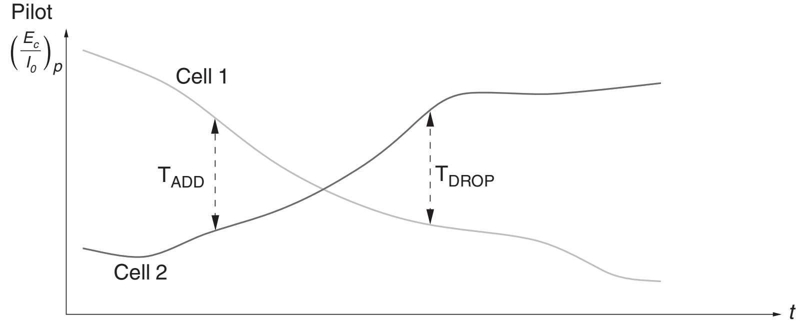


Figure 8.13 Soft handover in the WCDMA cellular system.

8.4.2 Fade margin improvement

Consider a hard handover scheme using the expected value of the signal strength from base stations (algorithm III of Example 8.1). The mobile terminal will always be connected to the closest base station, as illustrated in Figure 8.14. Hence, there is no diversity gain. Assuming shadow fading channels, the received SNR at the base station receiver can be written as follows:

$$\left(\frac{E_b}{N_0}\right)_i = \frac{GP_t}{L_pN}, \tag{8.18}$$

where P_t is the mobile terminal transmitted power, G is the shadow fading coefficient modeled as log-normal with σ dB standard deviation, L_p is the deterministic path loss, and N is the additive noise power. The outage probability, the probability of dropping a call before it is terminated by the mobile terminal or the base station when the mobile terminal is at the border of the cell, can be written as

$$P_{\text{out}} = \Pr\left(\frac{GMP_t}{L_pN_0} \leq \gamma_0 = \frac{P_t}{L_pN_0}\right) = Q\left(\frac{M_{\text{dB}}}{\sigma}\right) \leq p_{\text{out}}, \tag{8.19}$$

where p_{out} is the outage probability of interest, M is the required fade margin that ensures the outage probability of interest, γ_0 is the average received SNR at the border of the cell, and $Q(\cdot)$ is the Q-function. From (8.19), the required fade margin in hard handover is given by

$$M_{\text{dB}} = \sigma Q^{-1}(p_{\text{out}}) = \begin{cases} 1.2816\sigma, & p_{\text{out}} = 10\% \\ 2.3263\sigma, & p_{\text{out}} = 1\% \end{cases} \tag{8.20}$$

With soft handover, the mobile terminal is aware of the neighboring base stations and can connect to the best base station at any time, as illustrated in Figure 8.14. In this case the received-signal-energy-to-noise-power-spectral-density ratio from the mobile terminal will be the maximum of the received SNRs at the different base stations involved. Hence, soft handover in the uplink of DS-CDMA systems introduces selection diversity with an order that increases with the number of base stations involved in

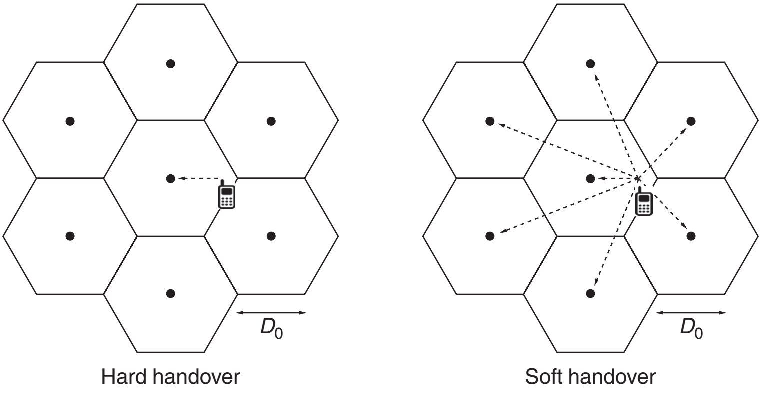


Figure 8.14 Hard and soft handover in wireless networks.

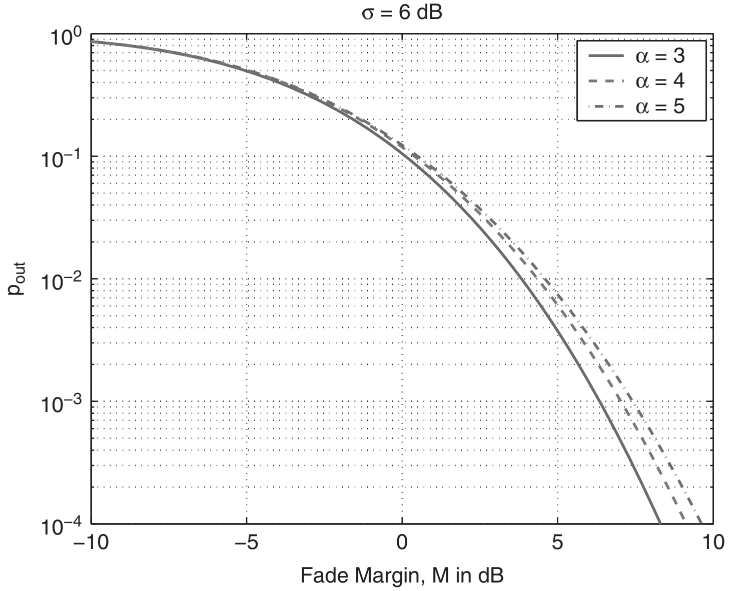


Figure 8.15 Outage probability of soft handover as a function of the required fade margin for the case of seven base stations within the active set.

the handover process. The outage probability when soft handover is employed can be written as

$$P_{\text{out}} = \Pr\left(\max\left\{\Gamma_1, \Gamma_2, \ldots, \Gamma_{N_c}\right\} \leq \gamma_0\right), \tag{8.21}$$

where N_c is the number of base stations involved in the soft handover process.

Assuming that the mobile terminal is at the border of the cell and considering the seven closest base stations as in Figure 8.14 with independent radio links, the outage probability is written as

$$p_{\text{out}} = \left[Q\left(\frac{M_{\text{dB}}}{\sigma}\right)\right]^3 \left[Q\left(\frac{M_{\text{dB}} - 3\alpha}{\sigma}\right)\right]^2 \left[Q\left(\frac{M_{\text{dB}} - 4.23\alpha}{\sigma}\right)\right]^2 \tag{8.22}$$

where α is the path loss exponent.

The outage probability of soft handover is illustrated in Figure 8.15 as a function of the fade margin for $\sigma = 6$ dB and for different values of the propagation path loss α. It is observed that the required fade margin is much lower compared to that required in hard handover. For instance, for an outage probability of 10% the required fade margin is about 0 dB, which is 7 dB better than that required by hard handover. For an outage probability of 1% the required fade margin is around 4 dB, which is 10 dB lower than that required by hard handover.

8.4.3 Effects of soft handover on DS-CDMA capacity

Referring to Chapter 5 and assuming perfect power control, the received-signal-energy-to-interference-plus-noise-power-spectral-density ratio of a given user is written as follows:

$$\left(\frac{E_b}{I_0}\right)_0 = \frac{W}{R}\frac{P}{(M-1)P + \sum_{b=1}^{B}\sum_{k=1}^{M}\left(\frac{G_{k,0}}{d_{k,0}^{\alpha}}\right)\left(\frac{d_{k,b}^{\alpha}}{G_{k,b}}\right)P + N_0 W},$$

where B is the number of base stations and M is the number of users per cell.

With a candidate set for soft handover of N_c cells, the mobile terminal is connected to the base station having the best link gain, i.e.,

$$\frac{G_{k,b}}{d_{k,b}^{\alpha}} = \max\left\{\frac{G_{k,1}}{d_{k,1}^{\alpha}}, \frac{G_{k,2}}{d_{k,2}^{\alpha}}, \ldots, \frac{G_{k,N_c}}{d_{k,N_c}^{\alpha}}\right\} > \frac{G_{k,0}}{d_{k,0}^{\alpha}}. \tag{8.23}$$

The capacity in this case can be approximated from the average received E_b/I_0 as follows:

$$\begin{aligned}\overline{\left(\frac{E_b}{I_0}\right)_i} &> \frac{W}{R}\frac{P}{(M-1)P + E\{I_{\text{inter}}\} + N_0 W} \\ &> \frac{W}{R}\frac{P}{\left(1+f_{N_c}\right)(M-1)P + N_0 W}.\end{aligned} \tag{8.24}$$

Here, due to the soft handover, we have an interference factor that depends on the number of cells in the candidate set for handover.

From the expression given in (8.24) we can solve for M and the uplink capacity of DS-CDMA cellular systems with soft handover can be written as

$$M \approx 1 + \frac{1}{F_{N_c}}\frac{W}{R}\left(\frac{1}{\xi_t} - \frac{1}{\xi_0}\right). \tag{8.25}$$

The interference factor, F_{N_c}, is a function of the number of base stations in the active set, the shadow fading parameters and the path loss exponent α. Table 8.2 gives some values for F_{N_c} for different values of N_c, σ and for $\alpha = 4$ [A. J. Viterbi et al., 1994]. It can be seen that the interference factor decreases with an increasing number of base stations within the active set. This decrease in the interference factor is more pronounced for large values of the standard deviation of the shadow fading channels. Higher values of the standard deviation cause more fluctuations in the radio links and hence a better diversity gain is achieved through soft handover.

Example 8.6: For a DS-CDMA system with a processing gain PG = 128 and a required signal-energy-to-interference-power-spectral-density ratio of $\xi_t = 7$ dB, the pole capacity with soft handover can be written as

$$M = 1 + \frac{1}{F_{N_c}}\frac{128}{5.01} \text{ users/cell.} \tag{8.26}$$

This is illustrated in Figure 8.16 for the different values of N_c taken from Table 8.2 and as a function of the standard deviation of the shadow fading channel.

Table 8.2 Interference factor of DS-CDMA systems with soft handover.

Standard deviation σ	$N_c = 1$	$N_c = 2$	$N_c = 3$	$N_c = 4$
0	1.44	1.44	1.44	1.44
2	1.48	1.43	1.43	1.43
4	1.67	1.47	1.45	1.45
6	2.13	1.56	1.49	1.49
8	3.38	1.77	1.57	1.55
10	7.17	2.28	1.75	1.66
12	20.8	3.62	2.17	1.91

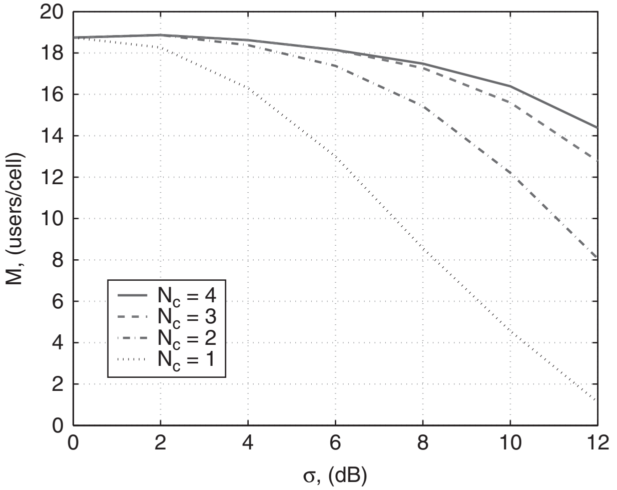


Figure 8.16 Impact of the number of soft handover cells.

On the downlink of DS-CDMA systems, a mobile terminal in soft handover requires a traffic channel from each base station involved in the soft handover process. Hence, the number of available traffic channels on the downlink decreases as the number of mobiles in soft handover increases and the interference within the involved cells increases.

The received signal energy to interference power spectral density ratio, E_b/I_0, of a certain mobile terminal m within the cell can be written as follows:

$$\left(\frac{E_b}{I_0}\right)_m = \frac{W}{R}\,\frac{\phi_m P_{1,m}}{\theta_m(1-\phi_m)P_{1,m} + \sum_{k=2}^{B} P_{k,m} + N_0 W} \geq \xi_t,$$

where θ_m (with $0 \leq \theta_m \leq 1$) is the orthogonality factor between the downlink users, B is the total number of base stations, ϕ_m is the fraction of power allocated to mobile terminal m, and $P_{k,m}$ is the received power from base station k at mobile terminal m.

When a certain mobile terminal k is in soft handover, with N_c base stations involved in the soft handover process, its combined E_b/I_0 at the rake receiver output can be written as:

$$\left(\frac{E_b}{I_0}\right)_k = \sum_{i=0}^{N_c-1}\left[\left(\frac{E_b}{I_0}\right)_k\right]_i,$$

where $\left[\left(\frac{E_b}{I_0}\right)_k\right]_i$ is the received E_b/I_0 from base station i.

Let us assume that there is only one mobile terminal in soft handover and the number of base stations involved in the soft handover process is $N_c = 2$. In this case the received E_b/I_0 of the mobile terminal in soft handover is written as

$$\begin{aligned}\left(\frac{E_b}{I_0}\right)_k &= \left[\left(\frac{E_b}{I_0}\right)_k\right]_0 + \left[\left(\frac{E_b}{I_0}\right)_k\right]_1 \\ &\approx \frac{W}{R}\frac{\phi_k P_{1,k}}{\theta\left(1-\frac{\phi_k}{2}\right)P_{1,k}+\sum_{i=2}^{B}P_{i,k}+N_0W} \geq \xi_t, \end{aligned} \tag{8.27}$$

where we have assumed that the total power allocated to this user is equally divided between the two base stations involved in the soft handover process. The other mobile terminals within the cell have a received E_b/I_0 given by

$$\left(\frac{E_b}{I_0}\right)_n = \frac{W}{R}\frac{\phi_n P_{1,n}}{\theta\left(1-\phi_n\right)P_{1,n}+\sum_{i=2}^{B}P_{i,n}+N_0W} \tag{8.28}$$

$$\geq \xi_t, \forall\, n \neq k \tag{8.29}$$

where ξ_t is the required E_b/I_0 threshold that ensures the QoS of the link.

The total number of active users within the cell is limited by the total transmit power of the base station. With ϕ_n denoting the fraction of transmit power allocated to mobile terminal n, the total number of mobile terminals in the downlink should satisfy the following inequality:

$$\phi_k + \sum_{n=1, n\neq k}^{M}\phi_n \leq 1. \tag{8.30}$$

The fraction of power needed for mobile terminal k, the one in soft handover, is obtained from (8.27) and can be written as

$$\phi_k \geq \frac{1}{\frac{\theta_k}{2}+\frac{W}{R}\frac{1}{\xi_t}}\left(\theta_k + f_{\mathrm{DL},k} + \frac{N_0W}{P_{1,k}}\right), \tag{8.31}$$

where $f_{\mathrm{DL},k}$ is as given in (5.75).

The fraction of power needed for the other mobile terminals within the cell is obtained from (8.29) and can be written as

$$\phi_n \geq \frac{1}{\theta_n+\frac{W}{R}\frac{1}{\xi_t}}\left(\theta_n + f_{\mathrm{DL},n} + \frac{N_0W}{P_{1,n}}\right), \quad \forall\, n \neq k. \tag{8.32}$$

Combining the inequality in (8.30) with those in (8.31) and (8.32) we get

$$\sum_{n=1}^{M}\left(\theta_n + f_{\mathrm{DL},n} + \frac{N_0 W}{PG_{1,n}}\right) - \phi_k \frac{\theta_k}{2} \leq \sum_{n=1}^{M} \phi_n \left(\theta_n + \frac{W}{R}\frac{1}{\xi_t}\right). \tag{8.33}$$

Solving for the total downlink base station transmission power we get

$$P \geq \frac{1}{\theta + \frac{W}{R}\frac{1}{\xi_t}} \frac{\sum_{m=1}^{M} \frac{N_0 W}{G_{1,m}}}{1 - M\frac{\theta + f_{\mathrm{DL}}}{\theta + \frac{W}{R}\frac{1}{\xi_t}} - \frac{\phi_k \theta_k}{2\left(\theta + \frac{W}{R}\frac{1}{\xi_t}\right)}}. \tag{8.34}$$

When the required transmit power approaches infinity, the downlink capacity approaches its maximum, the pole capacity. Hence, from (8.34) the downlink pole capacity when one mobile is in soft handover can be written as

$$M_{\mathrm{pole}} = \frac{\theta + \frac{W}{R}\frac{1}{\xi_t}}{\theta + f_{\mathrm{DL}}} - \frac{\phi_k \theta_k}{2\,(\theta + f_{\mathrm{DL}})}. \tag{8.35}$$

It can be seen that soft handover reduces the downlink capacity of DS-CDMA systems in comparison to the case without soft handover given in (5.81). The loss in capacity is a function of the orthogonality factor between the users within the cell and the multi-user interference. Compared to the case without soft handover, the capacity loss is then given by

$$F_h = \frac{\phi_k \theta_k}{2\,(\theta + f_{\mathrm{DL}})} \tag{8.36}$$

which is, in general, quite small since both ϕ_k and θ_k are parameters that are smaller than one.

In general, we may say that soft handover is an important mechanism that contributes to the QoS in DS-CDMA cellular systems. It increases the uplink capacity, improves the coverage, reduces the ping-pong effect and leads to a smooth mobility of users.

8.5 User association

We have discussed handover problems in the previous sections. Recall that the handover is basically the association between a mobile terminal and base stations while the terminal is on the move. The association problem may occur even when the terminal does not move. In this section, we will study some association issues when the terminal is stationary.

8.5.1 Load balancing

In wireless systems, the traffic load of a cell fluctuates heavily because terminals move around and the traffic can be generated at any time. Thus, neighboring cells may have quite different traffic loads temporarily. In this circumstance, the cell with the best signal quality is not always the cell that can provide the best service quality. Assume that you

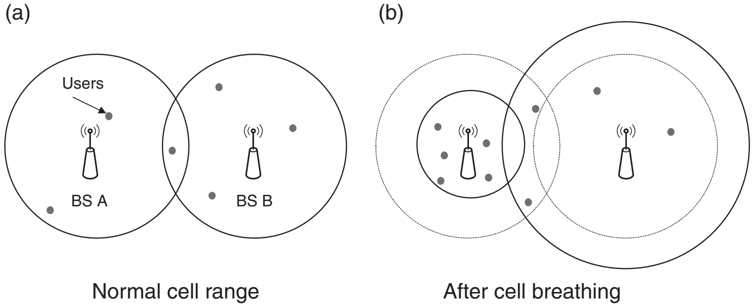


Figure 8.17 An illustration of cell breathing.

want to connect to cell *A* because the best SINR can be achieved with base station *A*. However, cell *A* already accommodates tens of users. There may be other cell(s), such as cell *B*, with a lower traffic load and still acceptable (though not the best) signal quality, which can eventually provide a better data rate. Then it would be beneficial to choose cell *B* rather than *A*. By associating with cell *B*, you can also benefit the users in cell *A* because they do not have to share the already insufficient resources with you. This is the concept of *load balancing* in wireless networks.

A simple way of implementing load balancing is to increase or decrease the output power of a cell-specific reference signal (which is also called the pilot channel or beacon signal) according to the traffic load. A congested cell decreases the signal power, and a cell with a light load increases it. Then mobile terminals "feel" that the light-loaded cell is closer and the congested cell is further away than they actually are. This is because the terminals estimate the signal quality of a base station by measuring the received reference signal power. As a consequence, cell sizes enlarge or shrink over time, like cells breathing in and out, as depicted in Figure 8.17. This technique is dubbed *cell breathing*.

Example 8.7: Assume a straight highway with 8 km length. Traffic intensity follows a location-dependent Poisson distribution with the intensity

$$\rho(x) = \begin{cases} x^2 & 0 \le x < 2 \\ 4 - \frac{x}{2} & 2 \le x \le 8 \\ 0 & \text{otherwise} \end{cases} . \tag{8.37}$$

We want to deploy two base stations to cover the highway. In order to balance the traffic load between the two base stations, where should the handover point be? Assume that the propagation loss is distance dependent and fading effects are neglected.

Solution: Let the traffic load of two cells be λ_1 and λ_2, respectively. Our objective is to find a handover point z which makes $\lambda_1 = \lambda_2$. Assume that z lies between 2 and 8, i.e. $2 \le z \le 8$. (You can also examine the case $0 \le z < 2$ with the same procedure explained below and verify that a feasible solution does not exist in this case.) The equation below should hold:

$$\lambda_1 = \int_0^2 x^2 \mathrm{d}x + \int_2^z (4 - \frac{x}{2})\mathrm{d}x = \lambda_2 = \int_z^8 (4 - \frac{x}{2})\mathrm{d}x. \tag{8.38}$$

This gives $z = 3.17$ km.

8.5.2 Association in heterogeneous networks

To refresh your memory, heterogeneous networks (HetNets) refers to a combination of base stations of different configurations. The fact that the base stations have different transmission powers makes association complicated. A fundamental problem in user association in HetNets is a mismatch between downlink and uplink in received signal quality. To simplify our discussion, we consider two base stations, one with a high power (e.g. macro cell) and the other with a low power (e.g. small cell). See Figure 8.18. In the downlink, the macro base station has much higher transmit power than the small base station. Thus, the macro base station will dominate the mid-point between the base stations, and the optimal downlink handover point is much closer to the small base station. However, the transmit power of a mobile terminal is invariant regardless of base station type, and thus the mid-point is the optimal uplink handover point. Traditionally in macro cellular systems, the association has been made based on the received downlink power. Simply applying the same rule to HetNets may lead to a serious problem in the uplink quality. Furthermore, since the macro base station accounts for most of the coverage area, small cells may serve only few users, making very little impact on the overall system performance.

A quick remedy to the mismatch is to introduce an artificial bias to the mobile terminal's association criterion [J. G. Andrews et al., 2012]. This is also known as *cell range extension* (CRE). Let s denote the index of the base station that you associate with. Among the set of candidate base stations Ψ, you need to choose s. Without the bias (and by momentarily neglecting the load balancing issue), your decision criterion will be

$$s = \arg\max_{i \in \Psi} \{P_{r,i}\}, \tag{8.39}$$

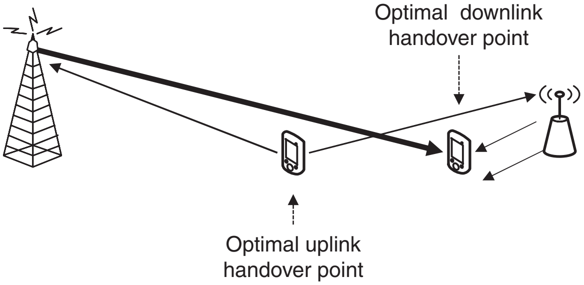


Figure 8.18 Mismatch in the downlink/uplink association in a HetNet.

where $P_{r,i}$ denotes received downlink power from base station i. By introducing the bias b_i to the cell i, the association changes to

$$s = \arg\max_{i \in \Psi}\{P_{r,i} + b_i\}. \tag{8.40}$$

Obviously, higher bias will force more users to associate with the small cells. While the bias can effectively solve the aforementioned problems, it may cause serious interference problems to the system. In fact, determination of optimal bias value is an open problem. You can guess that it depends on many parameters such as network deployment, traffic distribution, and service requirements.

Exercises

8.1 Consider a cellular system consisting of regular hexagonal cells with omni-directional antennas. The handover decision is based on a simple signal level criterion. The radio propagation experiences distance-dependent path loss (fourth power of distance) and log-normal shadow fading. Two handover decision schemes are considered:

I. Instantaneous decision: the terminal is associated with the strongest base station at any time.
II. Long-term decision: the terminal is associated with the base station of the highest long-term average signal power. Thus, it is always connected to the closest base station.

Consider a mobile at a point in cell 1 which is very close to the cell border. (This means that the mobile is almost at the same distance from adjacent base stations.) What is the difference between the median of received signal strengths with these handover decision schemes? Estimate the ratio of the two schemes for each of the four cases summarized in Table 8.3.

8.2 We will have a further look at the association problem in a HetNet. Consider a one-dimensional cellular system consisting of two base stations. Base station A is for macro coverage with a transmission power of 43 dBm, and base station B is for a small area with a transmission power of 37 dBm. The base stations use the same frequency band, and they are separated by the distance D. The base stations have the same antenna gain, and fading effects are ignored. The path loss is proportional to the fourth power of

Table 8.3 Cases for Exercise 8.1.

Case	Number of base stations involved in the handover	Shadow fading standard deviation
1	two	8 dB
2	two	3 dB
3	three	8 dB
4	three	3 dB

distance. Assume that a mobile terminal is moving straight from base station A to base station B.

a) Determine the location where the handover is made when the decision metric is downlink SIR.
b) Determine the location where the handover is made when the decision metric is uplink RSS.
c) You will find that the answers for (a) and (b) are not the same. This means that neither SIR nor RSS satisfies downlink and uplink simultaneously. How much worse does the downlink SIR get when uplink RSS is used as the metric? Also, what is the loss in uplink RSS when downlink SIR is used for the handover decision?

8.3 This exercise is a continuation of Example 8.4. In the example, we assumed $\lambda_h = \lambda_n$. Define the relative mobility α as

$$\alpha = \frac{\lambda_h}{\lambda_n + \lambda_h}. \tag{8.41}$$

Assume $N_{rsv} = 3$ and $\lambda_t = 6$. Calculate p_{block} and p_{drop} as functions of relative mobility. You may need a computer or calculator for the computation. Describe the impact of higher relative mobility on the performance of the cellular system.

8.4 Consider a cell serving a homogeneous delay-sensitive application (e.g. video telephony). New and handover sessions arrive at the cell following independent Poisson processes with intensities λ_n and λ_h, respectively. The sojourn time of each session follows an exponential distribution with the parameter μ. The cell can accommodate N_{tot} sessions in total. In order to prioritize handover sessions, we consider a probabilistic reservation scheme described as follows:

- Handover sessions are always accepted as long as there is capacity.
- When there are currently x sessions in the cell, a new session is accepted with the probability of $\frac{1}{x+1}$ $(x < N_{tot})$.

a) Draw a state transition diagram depicting this reservation scheme (for general N_{tot}).
b) Calculate the steady-state probability when $N_{tot} = 2$. Assume $\mu = 2\lambda_n = 2\lambda_h$.
c) Express handover dropping probability and new call blocking probability in terms of steady-state probabilities (for general N_{tot}).
d) Then obtain these values by using the same parameters as (b).

8.5 Cellular CDMA systems employ soft handover for both coverage and link reliability. A mobile receiver within an IS-95 system has been communicating for a while and receives the pilot signals shown in Figure 8.19. We assume that this mobile receiver is in soft handover and your task is to identify the set of cells that are in the active set of the mobile unit at time $\tau = \tau_0$. The handover drop time is 10 seconds and the handshaking time between the base station and the mobile can be neglected.

a) Going through each pilot signal, identify the status (in the active set, candidate set or neighbor set) of each pilot signal at time τ_0. Clearly explain your results.
b) Deduce the set of cells that are in the active set of the mobile unit at time τ_0.

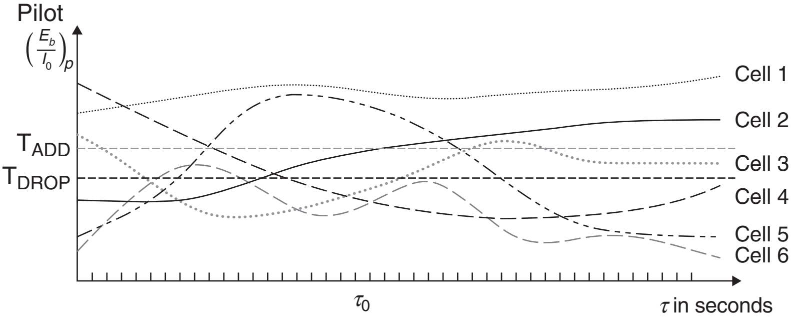


Figure 8.19 Channel variations.

References

J. G. Andrews, H. Claussen, M. Dohler, S. Rangan and M. C. Reed. 2012. Femtocells: past, present, and future. *IEEE Journal on Selected Areas in Communications*, 30(3), 497–508.

O. Grimlund and B. Gudmundson. 1991 (May 19–22). Handoff strategies in microcellular systems. Pages 505–510 of: *Proc. 41st IEEE Vehicular Technology Conference (VTC)*.

M. Gudmundson. 1991a (May). Analysis of handover algorithms. Pages 537–542 of: *41st IEEE Vehicular Technology Conference (VTC'91)*.

M. Gudmundson. 1991b. Correlation model for shadow fading in mobile radio systems. *Electronics Letters*, 27(23), 2145–2146.

I. Katzela and M. Naghshineh. 1996. Channel assignment schemes for cellular mobile telecommunication systems: a comprehensive survey. *IEEE Personal Communications*, 3(3), 10–31.

S. Sesia, M. Baker and I. Toufik. 2009. *LTE – The UMTS Long Term Evolution: From Theory to Practice*. Chichester: John Wiley & Sons.

A. J. Viterbi, A. M. Viterbi, K. S. Gilhousen and E. Zehavi. 1994. Soft handoff extends CDMA cell coverage and increases reverse link capacity. *IEEE Journal on Selected Areas in Communications*, 12(8), 1281–1288.

V. W.-S. Wong and V. C. M. Leung. 2000. Location management for next-generation personal communications networks. *IEEE Network*, 14(5), 18–24.

J. Zander and S.-L. Kim. 2001. *Radio Resource Management for Wireless Networks*. London: Artech House.

9 Energy-efficient design

9.1 Introduction

While semiconductor processing speed has been increasing exponentially, doubling almost every two years according to Moore's law, processor power consumption also continues to grow by 150% every two years [K. Lahiri et al., 2002]. By contrast, advances in battery technology have not kept pace, with capacity increasing at a modest rate of 10% every two years [K. Lahiri et al., 2002]. This leads to an increasingly large gap between power thirst and battery capacity. Information and communication technology (ICT) plays an important role in global greenhouse gas emissions since the amount of energy consumed by ICT is increasing dramatically to meet rapidly growing broadband mobile service requirements. For example, the power consumption for a macro base station can be 1400 watts and the corresponding energy costs can reach \$3200 per annum with 11 tons of CO_2 emissions. It has been shown that nowadays the total energy used by the infrastructure of cellular networks, wired networks and Internet takes up more than 2% of worldwide electrical energy consumption [GeSI, 2008]. The radio network itself adds up to 80% of an operator's entire energy consumption [EE Times, 2007]. In addition, this amount of energy is expected to increase rapidly in the coming years. Energy efficiency, therefore, is increasingly important for wireless mobile communications.

In this chapter we introduce some basic energy-efficient communication technologies. We start by studying node-level energy-efficient design, as improvements at the wireless node for energy-efficient radio transmission will translate into savings for the entire network. For an individual pair of wireless transceivers, the relation between power consumption, channel fading, path loss, modulation, coding, data rate and implementation factors are discussed thoroughly. To be specific, we will first analyze the energy consumption of different components of wireless transmitters. Then we will introduce the link energy efficiency metric that characterizes how efficiently energy is used in communication systems. Based on the energy efficiency metric, we will introduce how a communication pair can be designed in the most energy-efficient way. First, we will consider only radio transmission power consumption and show how the transmitter can be designed optimally to achieve the maximum energy efficiency. In practice, electronic circuits also consume a certain amount of operating power and this will significantly change the design of energy-efficient transmitters. We will study how this electronic circuit power consumption affects energy-efficient transmission. The optimal design that results in maximum total energy saving will be introduced.

In the latter half of the chapter, we will discuss network-level energy-efficient design. It is reported in the literature that currently over 80% of the power in mobile telecommunications is consumed in the radio access network, more specifically the base stations. Therefore we focus on energy-efficient design for radio access networks. Improvements in access network energy efficiency can be achieved in several ways. First, by optimization of individual cell sites, for example through the use of more efficient and load-adaptive radio resource management algorithms, such as discontinuous transmission and reception techniques. Secondly, by improved cell deployment strategies to effectively lower the number of sites required in the network to fulfill certain performance metrics such as coverage and spectral efficiency. In the improved deployment, classical macro-cellular topology can be complemented by low-power nodes, such as micro-, pico- or femtocells, to improve network energy efficiency. These techniques will be discussed in detail in the last few sections of this chapter.

9.2 Energy consumption in wireless networks

Devices in wireless networks have two main sources of energy consumption.

- Communication: energy spent by wireless communications. Communication energy depends on communication states, like SNR requirements, wireless channel (cell radius, fading, shadowing), network load and so on. In the following, this portion of energy consumption is termed transmission energy.
- Computation: energy spent by signal processing and computing to maintain the operation of the device. Computation energy depends on the hardware and software used for running the operating system, compressing data, coding and decoding, filtering and so forth. This portion of energy consumption is called circuit energy and is usually independent of communication states. Depending on the number of circuit components operating, the computation energy consumption may vary. For example, the energy consumption in sleep mode is usually much lower than that in idle mode.

The communication energy consumption dominates for long-distance communications such as those in macro cell communications. The computation energy consumption may dominate for short-distance communications, e.g. in femtocells or WiFi. For short-distance communications based on random access protocols, this is especially true when the network load is high, because a lot of protocol computation and signal processing takes place. Generally speaking, minimizing the total energy consumption requires minimizing the contributions of both communication and computation and finding the best tradeoff between them. We will discuss this later in this chapter. With the development of semiconductor technologies, which in general follow Moore's Law, the computation energy will get smaller and smaller. However, the communication energy is determined by the wireless channel and throughput requirement and does not follow

Moore's Law. Therefore the portion of energy consumed in communications in wireless networks is getting bigger and more important.

Energy is consumed by all layers of the implementations and network protocols, from radio frequency (RF) front to the operating system from physical and MAC layers to the whole network, and so on. Intuitively there are several ways to improve the energy efficiency of mobile devices. From the implementation perspective, low-power techniques can be applied in semiconductor manufacturing processes, circuit design, system architecture, operating system and application development. The consideration of energy efficiency will affect the whole design and implementation process. Our focus is on the communications side. One obvious solution is to reduce the amount of data that needs to be sent over the communication networks and one can use adaptive source coding and error control to reduce the number of information bits as much as possible. In the following we will focus on energy-efficient design of wireless networks so that the network energy consumption is minimized while the data is reliably delivered.

9.3 Energy-efficient transmission

In this section, we introduce energy-efficient techniques for point-to-point link communications. Energy efficiency quantifies how efficiently an energy resource is used. Therefore we define energy efficiency as the amount of information that can be reliably transmitted per unit energy consumption, that is,

$$u_e = \frac{Q}{E}, \tag{9.1}$$

where Q is the number of information bits successfully delivered and E is the corresponding energy consumption. The unit of the energy efficiency is bits per joule. Q is usually a fixed value, for example a packet size, that depends on the source coding. Dividing both numerator and denominator by the time used, an equivalent energy efficiency definition is

$$u_e = \frac{S}{P}, \tag{9.2}$$

where S is the average throughput and P the average power consumption. The goal of energy-efficient transmission is to maximize the energy efficiency u_e. Note that by maximizing u_e, the amount of energy needed to send the Q bits is minimized. This can be seen by observing the following equivalent designs,

$$\max u_e \Leftrightarrow \max \frac{Q}{E} \overset{\text{Given } Q}{\Leftrightarrow} \min E. \tag{9.3}$$

So the goal of energy-efficient design is to minimize the energy consumption for delivering each information bit, or equivalently, maximize the number of information bits reliably transmitted per unit energy consumption.

9.3.1 Ideal transmission

With capacity-approaching channel codes, such as turbo codes, the reliable data rate of an AWGN channel is given by

$$S = W\log_2(1+\frac{Pg}{WN_0\theta}), \tag{9.4}$$

where P is the transmission power, g the channel power gain, W the signal bandwidth, N_0 the noise spectral density, and $\theta \geq 1$ the SINR gap that defines the gap between the channel capacity and a practical coding and modulation scheme. If the channel capacity is achieved, $\theta = 1$. The time to deliver one bit is t, which equals

$$t = \frac{1}{S} = \frac{1}{W\log_2(1+\frac{Pg}{WN_0\theta})}, \tag{9.5}$$

and the power–time relationship is

$$P = \left(2^{\frac{1}{Wt}} - 1\right) WN_0\theta/g. \tag{9.6}$$

Thus the relationship between the energy and time used in sending this one bit is

$$E = pt = \left(2^{\frac{1}{Wt}} - 1\right) WtN_0\theta/g, \tag{9.7}$$

and the energy efficiency is

$$u_e = \frac{1}{E} = \frac{g}{\left(2^{\frac{1}{Wt}} - 1\right) WtN_0\theta}. \tag{9.8}$$

The energy is monotonically decreasing and convex in the transmission time t, as illustrated in Figure 9.1. Achieving the highest energy efficiency $u_e = \frac{1}{E}$ is the same as minimizing the energy consumption E. We can see that the energy is minimized when t approaches infinity, i.e. using an infinite amount of time to send the one bit. In this case, an infinitely small amount of power will be used according to (9.6). To ensure reliable transmission, we need to use a coding scheme with an infinitely small coding rate. The minimum energy needed to send one bit is

$$E_{min} = \lim_{t->\infty} \left(2^{\frac{1}{Wt}} - 1\right) WtN_0\theta/g = \frac{N_0\theta\ln 2}{g}. \tag{9.9}$$

Therefore the upper bound of energy efficiency is

$$u_e \leq \frac{1}{E_{min}} = \frac{g}{N_0\theta\ln 2}. \tag{9.10}$$

We can also look at the energy consumption from the receiver perspective:

$$\frac{Eg}{N_0} \geq \frac{E_{min}g}{N_0} = \theta\ln 2 \geq \ln 2 = -1.59 \text{ dB}, \tag{9.11}$$

indicating that the minimum received signal energy in total can be 1.59 dB smaller than the noise spectral density. A transmission that achieves this goal will also minimize its transmission energy consumption.

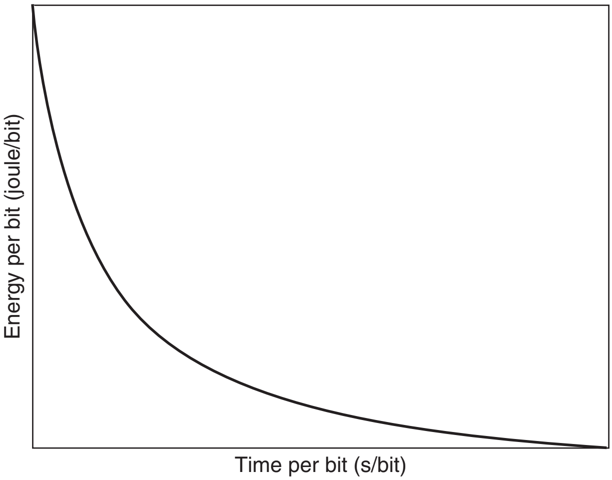


Figure 9.1 Energy consumption and time to send one bit.

9.3.2 Energy-efficient transmission in practice

As discussed in Section 9.2, in addition to transmission energy, wireless devices also consume circuit energy. When circuit energy is taken into account, the method of transmitting with the longest duration is no longer the best since circuit energy consumption increases with transmission duration. In this case, the overall energy for transmitting one bit over an AWGN channel turns out to be

$$E = \left(\frac{P}{\zeta} + P_c\right) t = \frac{1}{\zeta}\left(2^{\frac{1}{Wt}} - 1\right) WN_0\theta t/g + P_c t, \tag{9.12}$$

where $\zeta \in [0,1]$ is the power amplifier efficiency and depends on the design and implementation of the transmitter. P_c is the average circuit power, including all electronic power consumption, except transmission power, for reliable data transmission. The energy consumption is no longer monotonically decreasing in the time t. Figure 9.2 shows the relationships of different components of energy consumption and the time used to send one bit. While the transmission energy decreases with t, the circuit energy grows. Therefore energy-efficient transmission should find the optimal tradeoff between transmission energy and circuit energy consumptions. Since the tradeoff is determined by the time for sending one bit, we need to find the optimal time, or equivalently, its inverse, the optimal reliable data rate, which is determined by the link adaptation that includes adapting power, and the MCS.

To transmit an arbitrary number of bits, we want to find the optimal link adaptation to maximize its energy efficiency,

$$\max_S u_e(S) = \max_S \frac{S}{\frac{P}{\zeta} + P_c} = \max_S \frac{S}{\left(2^{\frac{S}{W}} - 1\right)\frac{WN_0\theta}{g\zeta} + P_c}. \tag{9.13}$$

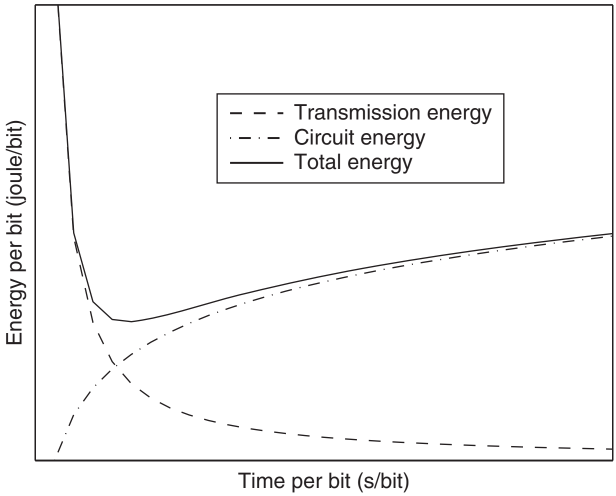


Figure 9.2 Relationship between energy consumption and time used to send one bit.

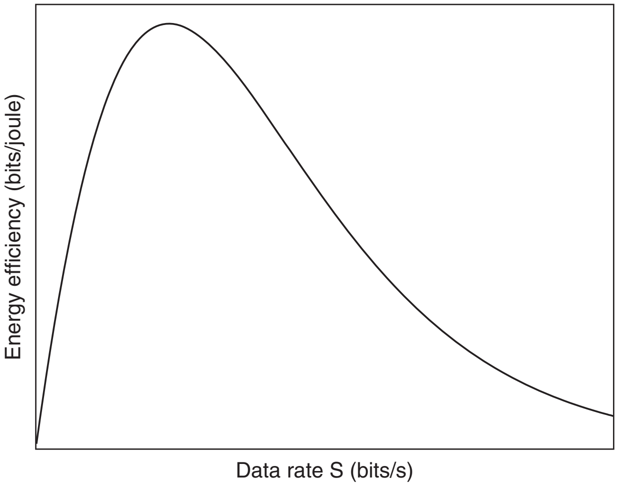


Figure 9.3 Relationship between energy efficiency and data rate.

In general, it has been shown [G. W. Miao et al., 2010] that energy efficiency u_e is strictly quasi-concave in S, as illustrated in Figure 9.3, and has a local maximum.* A strictly quasi-concave function has a unique maximum if it has a local maximum. Therefore there is a unique optimal S that maximizes the energy efficiency, which is what we need for energy-efficient transmission. The optimal S can be found by setting the first-order derivative of u_e to be zero. The necessary and sufficient condition for a data rate to be globally optimal is given below. For general wireless transmissions, the following theorem summarizes the optimal energy-efficient setting.

* The reader is referred to [M. S. Bazaraa et al., 2013] for more information about quasi-concave programming.

THEOREM 9.1 *There exists a unique globally optimal transmission data rate to maximize the energy efficiency and it is given by [G. W. Miao et al., 2008]*

$$S^* = \frac{P_c + P(S^*)}{P'(S^*)}, \tag{9.14}$$

where $P(S)$ is the transmission power needed to achieve rate S. $P'(S)$ is its first-order derivative. In addition, the energy efficiency is bounded above by

$$u_e \leq \frac{1}{P'(0)}. \tag{9.15}$$

When $P(S) = \left(2^{\frac{S}{W}} - 1\right) \frac{WN_0\theta}{g\zeta}$,

$$u_e \leq \frac{g\zeta}{N_0\theta \ln 2}. \tag{9.16}$$

The optimal S can be found easily using some numeric algorithms like binary search [G. W. Miao et al., 2010]. The basic idea of binary search is to first identify two rates, S_1 and S_2, such that $S_1 < S^* < S_2$. Then we split the range $[S_1, S_2]$ into two smaller ranges, $[S_1, \frac{S_1+S_2}{2}]$ and $[\frac{S_1+S_2}{2}, S_2]$, and determine which range S^* falls in. Update S_1 and S_2 to be the new smaller range and iterate the process until the desired accuracy is achieved, for example when $|S_1 - S_2| < \delta$. In implementation, the achievable rate has discrete values and the one closest to S^* can be used.

In the following, we investigate some basic properties of energy-efficient link adaptation. Propositions 9.1, 9.2 and 9.3 summarize the impact of channel gain, circuit power and signal bandwidth on the optimal energy-efficient transmission.

PROPOSITION 9.1 *Both the optimal data rate S^* and the maximum energy efficiency, $u_e(S^*)$, increase with channel gain g [G. W. Miao et al., 2010].*

PROPOSITION 9.2 *The optimal data rate S^* increases with circuit power, while the energy efficiency, $u_e(S^*)$, decreases with it. With zero circuit power, the highest energy efficiency $\frac{g\zeta}{N_0\theta \ln 2}$ is obtained by transmitting with infinitely small data rate [G. W. Miao et al., 2010].*

PROPOSITION 9.3 *The energy efficiency $u_e(S^*)$ increases with signal bandwidth.*

Propositions 9.1, 9.2 and 9.3 suggest three ways to improve energy efficiency: increasing channel power gain, reducing circuit power, and allocating more bandwidth. The channel power gain can be improved by moving the transmitter and receiver closer and the circuit power can be improved using low-power circuit design technologies.

The signal bandwidth is controlled by the medium access control scheme. For example, in a cellular network the signal bandwidth of each user is scheduled by the base station. Based on Proposition 9.3, each user desires more bandwidth for higher energy efficiency. However, the entire system bandwidth cannot be allocated exclusively to one user as this would adversely affect the energy efficiency of other users as well as that of the overall network. This is illustrated in Figure 9.4, where a two-user OFDMA network with ten subcarriers is considered. We can see that there is

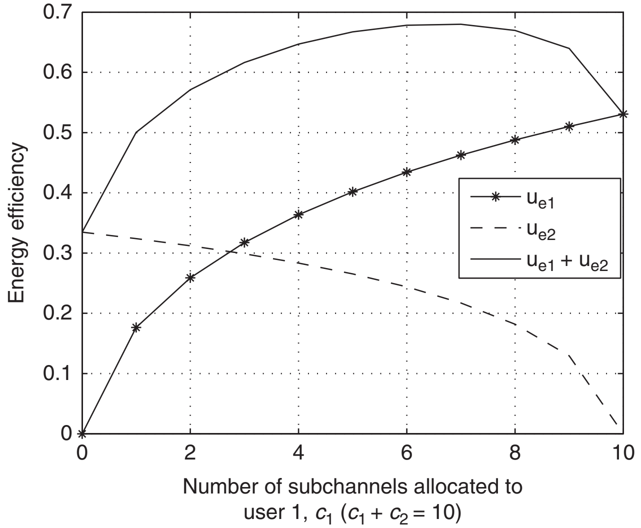


Figure 9.4 Energy efficiency of a two-user OFDMA network.

an optimal subcarrier allocation that maximizes the network energy efficiency. Hence, the subcarrier or bandwidth allocation is critical in determining the overall network energy efficiency. Energy-efficient scheduling algorithms similar to those introduced in Chapter 4 are needed to improve the energy efficiency of the overall network [G. W. Miao et al., 2008, 2012].

Example 9.1: An ideal wireless transmitter needs to send 100 bits in the buffer over an AWGN channel. It consumes a circuit power of 1 W. All parameters, if not specified, are assumed to have a value of 1.

(1) How fast should it transmit such that it is the most energy efficient? What is the minimum amount of energy needed to send these bits?
(2) Repeat (1) assuming the circuit power is zero.
(3) Repeat (1) assuming the circuit power is infinite.

Solution: (1) The energy efficiency is

$$u_e = \frac{S}{\frac{P}{\zeta} + P_c} = \frac{\log_2(1 + P/1)}{P + 1},$$

which is maximized when $P = 1.72$ W. The data rate is

$$S^* = \log_2(1 + 1.72) = 1.44 \text{ bits/s}.$$

The time is

$$t^* = \frac{100}{S} = 100/1.44 = 69.27 \text{ s}.$$

Correspondingly,

$$u_e^* = 0.53 \text{ b/J}.$$

The minimum energy needed is

$$E^* = \frac{100}{u_e^*} = 188.7 \text{ J}.$$

(2) When $P_c = 0$,

$$u_e = \frac{S}{\frac{P}{\zeta}} = \frac{\log_2(1+P/1)}{P},$$

which is maximized when $P \to 0^+$. Correspondingly,

$$S^* \to 0^+$$

and

$$t^* \to \infty.$$

The minimum energy needed is

$$E^* = \frac{100}{u_e^*} = \frac{100}{\lim_{P\to 0^+} \frac{\log_2(1+P/1)}{P}} = 100/1.44 = 69.4 \text{ J}.$$

(3) When $P_c \to \infty$,

$$u_e = \frac{S}{\frac{P}{\zeta}} = \frac{\log_2(1+P)}{P+\infty},$$

which is maximized when $P = \infty$. The energy efficiency is $u_e^* = 0$ and an infinite amount of energy is needed to send 100 bits.

9.3.3 Energy-efficient transmission in frequency-selective channels

In wideband applications, different frequency bands usually experience different levels of fading. Current communication systems design deals with frequency selectivity through subdividing the bandwidth into small segments, where the channel can be assumed to be flat. So for an ideal channel orthogonalization technology such as MIMO (multiple input, multiple output) or OFDM, the channel may be divided into K subchannels, each experiencing flat fading. We will consider static channels to gain insights; the results for time-varying channels can be derived similarly. Assume that K subchannels are used for transmission, each with a different channel gain. An example of this scenario is OFDM transmission over frequency-selective channels and MIMO transmission over several spatial channels. Another example is the downlink transmission of a base station serving multiple users at the same time using OFDMA or SDMA. After scheduling and subchannel assignment, the base station needs to determine power, modulation order and coding on all the subchannels. Define the data rate on subchannel i as r_i and the vector of data rates on all subchannels as

$$\mathbf{R} = [r_1, r_2, \ldots, r_K]^T, \tag{9.17}$$

where $[]^T$ is the transpose of a vector. The data rate vector, $\mathbf{R}$, depends on the channel state, coding and power allocation. Correspondingly, the overall data rate is

$$R = \sum_{i=1}^{K} r_i. \tag{9.18}$$

Define the total transmission power to be $P(\mathbf{R})$, which is assumed to be strictly convex and monotonically increasing in $\mathbf{R}$ with $P(\mathbf{0}) = 0$, where $\mathbf{0} = [0, 0, \ldots, 0]^T$. $P(\mathbf{R})$ captures all the power consumption that is dependent on $\mathbf{R}$. In practice, we need to find $P(\mathbf{R})$ for a transmission system so that the energy-efficient link adaptation technique can be applied to improve its energy efficiency.

An example of $P(\mathbf{R})$ is given as follows; more examples can be found in [G. W. Miao and G. Song, 2015]. The output SNR on subchannel i is

$$\Gamma_i = \frac{P_i g_i}{N_o W}, \tag{9.19}$$

and the achievable data transmission rate r_i is

$$r_i = W \log_2 \left(1 + \frac{\Gamma_i}{\theta}\right). \tag{9.20}$$

The overall transmit power is

$$P(\mathbf{R}) = \frac{\sum_{i=1}^{K} P_i}{\zeta} = \sum_{i=1}^{K} (2^{\frac{r_i}{W}} - 1) \frac{N_o W \theta}{g_i \zeta}. \tag{9.21}$$

We can see that $P(\mathbf{R})$ is strictly convex and monotonically increasing in $\mathbf{R}$.

The overall power consumption given a data rate vector is

$$P(\mathbf{R}) = P_c + P(\mathbf{R}). \tag{9.22}$$

The system energy efficiency is

$$u_e(\mathbf{R}) = \frac{R}{P_c + P(\mathbf{R})}. \tag{9.23}$$

The optimal energy-efficient link adaptation achieves maximum energy efficiency, i.e.

$$\mathbf{R}^* = \arg\max_{\mathbf{R}} u_e(\mathbf{R}) = \arg\max_{\mathbf{R}} \frac{R}{P_c + P(\mathbf{R})}. \tag{9.24}$$

Note that if we fix the overall transmit power, the objective of Equation (9.24) is equivalent to maximizing the overall throughput. However, besides adapting the power distributions on all subchannels, the overall transmit power can also be adapted according to the states of all subchannels to maximize the energy efficiency. Hence, the solution to Equation (9.24) is in general different from existing power allocation schemes that maximize throughput with power constraints, e.g. the water-filling solution for OFDMA in (4.38).

In [G. W. Miao et al., 2010] it is shown that if $P(\mathbf{R})$ is strictly convex in $\mathbf{R}$, $u_e(\mathbf{R})$ is strictly quasi-concave and has a local maximum. For strictly quasi-concave functions, if a local maximum exists, it is also globally optimal. Hence, a unique globally optimal transmission rate vector always exists and its characteristics are summarized in Theorem 9.2.

THEOREM 9.2 *[G. W. Miao et al., 2010] If $P(\mathbf{R})$ is strictly convex, there exists a unique globally optimal transmission data rate vector $\mathbf{R}^* = [r_1^*, r_2^*, \ldots, r_K^*]^T$ for (9.24), where r_i^* is given by*

(i) when $\frac{P_c+P(\mathbf{R}_i^{(0)})}{R_i^{(0)}} \geq \left.\frac{\partial P(\mathbf{R})}{\partial r_i}\right|_{\mathbf{R}=\mathbf{R}_i^{(0)}}$, $\left.\frac{\partial u_e(\mathbf{R})}{\partial r_i}\right|_{\mathbf{R}=\mathbf{R}^*} = 0$, *i.e.* $\frac{1}{\frac{\partial P(\mathbf{R}^*)}{\partial r_i^*}} = \frac{R^*}{P_c+P(\mathbf{R}^*)} = u_e(\mathbf{R}^*)$;

(ii) when $\frac{P_c+P(\mathbf{R}_i^{(0)})}{R_i^{(0)}} < \left.\frac{\partial P(\mathbf{R})}{\partial r_i}\right|_{\mathbf{R}=\mathbf{R}_i^{(0)}}$, $r_i^* = 0$,

where $\mathbf{R}_i^{(0)} = [r_1^*, r_2^*, \ldots, r_{i-1}^*, 0, r_{i+1}^*, \ldots, r_K^*]$ *and* $R_i^{(0)} = \sum_{j \neq i} r_j^*$, *i.e. the overall data rate on all other subchannels except i.*

Theorem 9.2 has clear physical insights. $P_c + P(\mathbf{R}_i^{(0)})$ is the power consumption of both circuit and all other subchannels when subchannel i is not used. $\frac{P_c+P(\mathbf{R}_i^{(0)})}{R_i^{(0)}}$ is the per-bit energy consumption when subchannel i is not used and the overall per-bit energy consumption needs to be minimized for energy-efficient communications. $\left.\frac{\partial P(\mathbf{R})}{\partial r_i}\right|_{\mathbf{R}=\mathbf{R}_i^{(0)}}$ is the per-bit energy consumption transmitting at an infinitely small data rate on subchannel i conditioned on the optimal status of all other subchannels. Hence, subchannel i should not transmit anything when $\frac{P_c+P(\mathbf{R}_i^{(0)})}{R_i^{(0)}} < \left.\frac{\partial P(\mathbf{R})}{\partial r_i}\right|_{\mathbf{R}=\mathbf{R}_i^{(0)}}$. Otherwise, there should be a tradeoff between the desired data rate on subchannel i and the incurred power consumption. The tradeoff closely depends on the power consumption of both circuits and transmission on all other subchannels and can be found through the unique zero derivative of $u_e(\mathbf{R})$ with respect to r_i.

To further understand Theorem 9.2, we consider an example where each subchannel achieves the Shannon capacity and the transmit power on each subchannel is given in (9.21) with $\theta = 0$ dB and $\zeta = 1$. The overall transmit power is

$$P(\mathbf{R}) = \sum_{k=1}^{K} (2^{\frac{r_k}{W}} - 1)\frac{N_o W}{g_k}. \tag{9.25}$$

According to condition (i) of Theorem 9.2, when $r_k > 0$, we have

$$\frac{1}{\frac{\partial P(\mathbf{R})}{\partial r_k}} = \frac{1}{2^{\frac{r_k}{W}}\frac{N_o}{g_k}} = u_e(\mathbf{R}^*). \tag{9.26}$$

Hence, the transmit power on subchannel k is

$$P_n = (2^{\frac{r_k}{W}} - 1)\frac{N_o W}{g_k} = \frac{W}{u_e(\mathbf{R}^*)} - \frac{N_o W}{g_k}, \tag{9.27}$$

which is a water-filling allocation with water level $\frac{W}{u_e(\mathbf{R}^*)}$, as illustrated in Figure 9.5. Since the water level is determined by the optimal energy efficiency, the scheme can be called dynamic energy-efficient water-filling. Note that while the absolute value of power allocation is determined by the maximum energy efficiency $u_e(\mathbf{R}^*)$, which relies

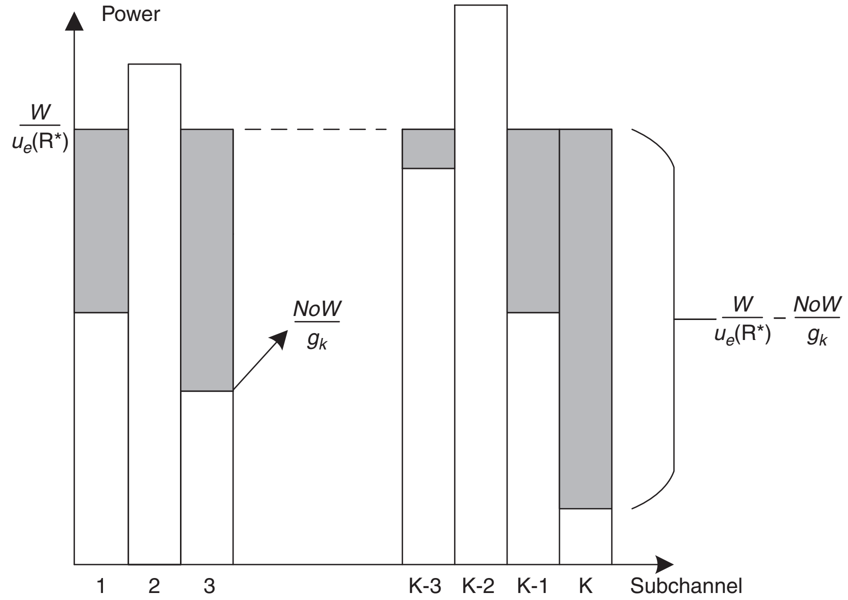


Figure 9.5 Dynamic energy-efficient water-filling power allocation.

on both the circuit power and channel state, the relative differences of power allocation on different subchannels depend only on the channel gains on those subchannels.

When the link adaptation has physical requirements, e.g. peak or average power limits, data rate requirement and so on, it should achieve the highest energy efficiency while meeting these requirements. If the globally optimal link adaptation meets these requirements, it can be used. Otherwise, the transmission should be adapted to the boundary conditions. This is illustrated in Figure 9.6, in which a user has a minimum data rate requirement R_{min} and maximum data rate limit R_{max}. The optimal transmission rates in the three cases in Figure 9.6 are R_{max}, R^* and R_{min} respectively, where R^* is determined by Theorem 9.2.

Theorem 9.2 provides the necessary and sufficient conditions for a rate vector to be the unique and globally optimum one. It is usually difficult to directly solve the joint nonlinear equations according to Theorem 9.2. Quasi-concave programs share some important properties with concave programs and many standard methods for concave programs can be used to solve the energy-efficiency maximization problem. Numerical algorithms with hill-climbing procedures can be applied to search for the optimal $\mathbf{R}$ for maximizing $u_e(\mathbf{R})$. Many known nonlinear optimization techniques are available for this purpose, for example the classical subgradient methods and Newton's method for unconstrained or the interior point method for the constrained energy-efficient link adaptation. The details of these numerical algorithms are beyond the scope of this text and two examples can be found in [G. W. Miao et al., 2010; C. Isheden and G. P. Fettweis, 2011], where the first is based on the method of steepest ascent and the second is based on Dinkelbach's algorithm [W. Dinkelbach, 1967], which is a variation of Newton's method. While these iterative algorithms have different convergence

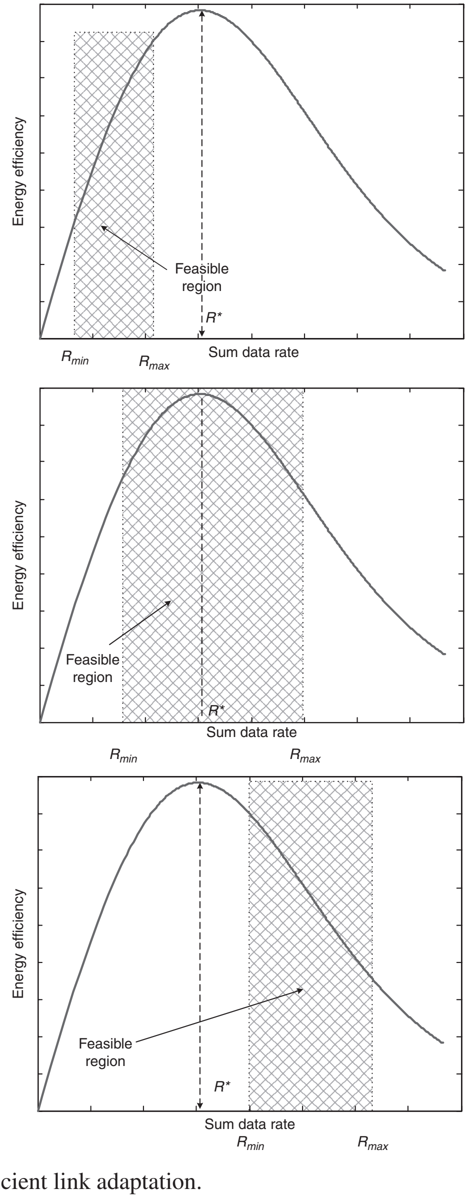


Figure 9.6 Constrained energy-efficient link adaptation.

speeds, they are all iterative and consume a lot of computing power. Low-complexity algorithms, e.g. closed-form ones, that perform close to the optimal ones are essential to reduce the computation power consumption. The reader interested in closed-form link adaptation techniques may refer to [G. W. Miao et al., 2012] for more discussions.

Example 9.2: A wireless transmitter uses two-subcarrier OFDM and on each subcarrier, adaptive M-ary quadrature amplitude modulation (M-QAM) is applied, whose BER performance on each subcarrier is approximated by

$$p_e = 0.2 \exp \frac{1.5\Gamma_i}{2^{Kr_i} - 1}, \tag{9.28}$$

where K is a constant and i the subcarrier index. It consumes P_1 circuit power in transmission mode and P_2 circuit power in sleep mode. The transmitter needs to send a packet of Q bits once every T seconds. After sending all bits, it goes to the sleep mode (see Figure 9.7). Determine the energy-efficient link adaptation for this transmitter with a BER requirement p_e on both subcarriers.

Solution: Let P_3 denote the power consumption for reliable data communications; according to (9.28),

$$P_3 = \ln(5p_e) \sum_i (2^{Kr_i} - 1)2/3 \frac{N_0 W}{g_i}. \tag{9.29}$$

The transmitter energy efficiency is

$$\begin{aligned} u_e &= \frac{Q}{(P_2 + P_3)t + P_1(T - t)} \\ &= \frac{Q}{\left(P_2 + \sum_i \ln(5p_e)(2^{Kr_i} - 1)2/3\frac{N_0 W}{g_i}\right)\frac{Q}{\sum_i r_i} + P_1(T - \frac{Q}{\sum r_i})} \\ &= \frac{\sum r_i}{\left[\ln(5p_e)\sum_i (2^{Kr_i} - 1)2/3\frac{N_0 W}{g_i} + P_1 T(\sum_i r_i)/Q\right] + (P_2 - P_1)}. \end{aligned} \tag{9.30}$$

Therefore,

$$P(r) = \left[\ln(5p_e)\sum_i (2^{Kr_i} - 1)2/3\frac{N_0 W}{g_i} + P_1 T \sum_i r_i/Q\right] \tag{9.31}$$

and

$$P_c = P_2 - P_1. \tag{9.32}$$

Then Theorem 9.2 can be used to obtain the optimal data rate vector $\mathbf{R}^*$. If $R^* < \frac{Q}{T}$, $\frac{Q}{T}$ should be the optimal transmission data rate.

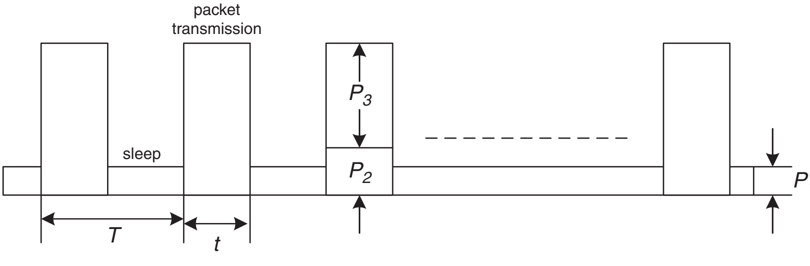


Figure 9.7 Illustration of a wireless transmitter.

9.4 Tradeoff in network resource utilization

Both spectral efficiency (SE) and energy efficiency (EE) emphasize communication quality in the sense that successful data transmission will help improve both metrics. From this perspective EE also desires SE improvement and vice versa. On the other hand, energy and spectrum are two independent fundamental resources and both are needed to meet QoS requirements. Tradeoff is always there and depends on network choice. Energy and spectral efficiency are equally important and there is no clear advantage of one metric over the other. Which metric is more desired will depend on network needs. In this section, we discuss the fundamental tradeoffs in energy and spectrum efficiency in wireless networks.

9.4.1 Energy and spectral efficiency in interference-free channels

In this section, we investigate the tradeoff between spectral and energy efficiency in a single-user system without interference.

When there is only one user in the network, the spectral efficiency of the system is

$$u_s = \log_2\left(1+\frac{Pg}{WN_0}\right). \tag{9.33}$$

Assuming first an ideal transmitter with zero circuit power, the energy efficiency of the system is

$$u_e = \frac{W\log_2\left(1+\frac{Pg}{WN_0}\right)}{P}. \tag{9.34}$$

Therefore the spectral and energy efficiency can be characterized by the following equation,

$$u_e = \frac{u_s g}{(2^{u_s}-1)N_0}. \tag{9.35}$$

This relationship is illustrated in Figure 9.8 and u_e is strictly decreasing in u_s.

In practice, the circuit power is not zero and the energy efficiency is

$$u_e = \frac{W\log_2\left(1+\frac{Pg}{WN_0}\right)}{P+P_c}. \tag{9.36}$$

Therefore we have the tradeoff relationship

$$u_e = \frac{u_s}{(2^{u_s}-1)N_0/g+P_c/W}, \tag{9.37}$$

which is illustrated in Figure 9.9, where $P_{c1}=0$.

9.4.2 Energy and spectral efficiency in interference channels

In this section, we consider a multi-user system where different users interfere with each other. To facilitate analysis and get insights, consider a symmetric network. There

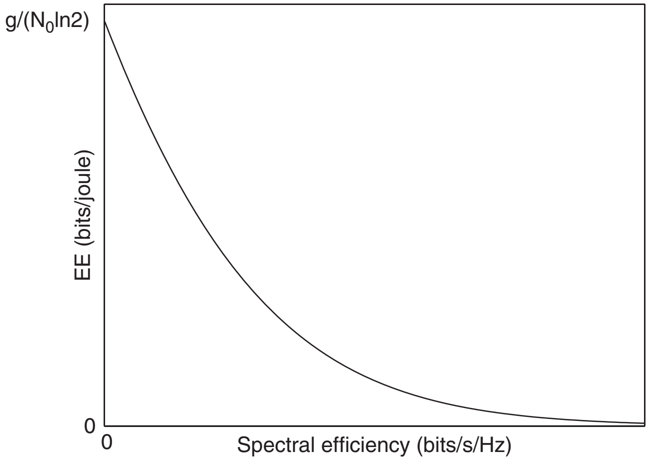


Figure 9.8 Tradeoff between EE and SE with zero circuit power.

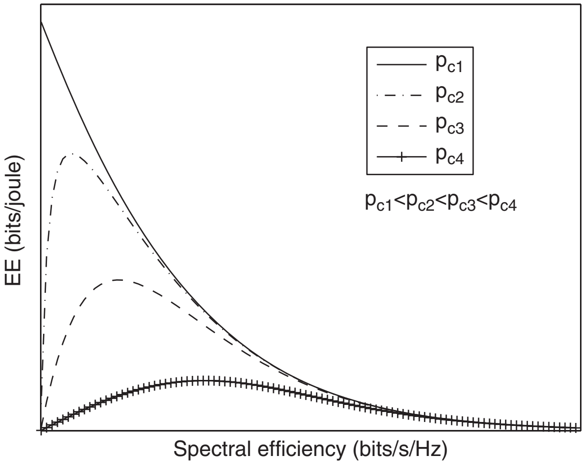


Figure 9.9 Tradeoff between EE and SE in practice.

are N users, all experiencing the same channel power gain g. All interference channels have the same power gain $\widetilde{g}$. Define the network coupling factor

$$\phi = \frac{\widetilde{g}}{g}, \tag{9.38}$$

which is a network property and characterizes how different links interfere with each other compared against their own links. Higher ϕ represents a heavier interfering scenario. Assume the transmission power of all users to be P.

The overall network energy efficiency will be

$$\begin{aligned} u_e(P) &= \sum_{n=1}^{N} \frac{W\log_2\left(1+\frac{Pg}{\sum_{i,i\neq n} P\tilde{g}+N_0W}\right)}{P+P_c} \\ &= \frac{NW\log_2\left(1+\frac{P}{(N-1)\phi P+\frac{N_0W}{g}}\right)}{P+P_c}, \end{aligned} \tag{9.39}$$

and the network spectral efficiency will be

$$u_s(P) = N\log_2\left(1+\frac{P}{(N-1)\phi P+\frac{N_0W}{g}}\right). \tag{9.40}$$

The following tradeoff relationship can be readily derived:

$$u_e = \frac{Wu_s}{P_c+\frac{AN_0W}{g[1-A(N-1)\phi]}}, \tag{9.41}$$

where $A = 2^{\frac{u_s}{N}} - 1$.

Assume first the power of each user is chosen to maximize its own spectral efficiency. As u_s is strictly increasing in P, the upper bound of the network spectral efficiency is

$$u_s^* = \lim_{P\to\infty} u_s(P) = N\log_2\left(1+\frac{1}{(N-1)\phi}\right) \tag{9.42}$$

with the corresponding energy efficiency $u_e = \lim_{P\to\infty} u_e(P) = 0$, which is completely energy inefficient.

Second, if the power of each user is chosen to maximize its own energy efficiency and

$$P^* = \arg\max_P u_e(P),$$

the network energy efficiency is

$$u_e^* = u_e(P^*)$$

with the corresponding spectral efficiency $u_s = u_s(P^*)$. Hence, the largest possible spectral efficiency penalty, i.e. loss in spectral efficiency, by using energy-efficient power optimization is

$$\triangle u_s = u_e^* - u_s = N\log_2(1+\frac{1}{(N-1)\phi}) - u_s(P^*). \tag{9.43}$$

In an interference-free scenario, i.e. $N = 1$ or $\phi = 0$, the penalty is infinite. Otherwise, whenever interference exists, it is bounded.

To further understand the tradeoff, Figure 9.10 (from [G. W. Miao et al., 2011]) illustrates a case where two users are transmitting using the same power and interfering with each other. The curves with markers illustrate the relationships between the transmission power P and the spectral efficiency u_s when the network has different coupling factors ϕ, while those without markers give the corresponding energy efficiency u_e. When $\phi = 0$, an arbitrarily high spectral efficiency can be achieved

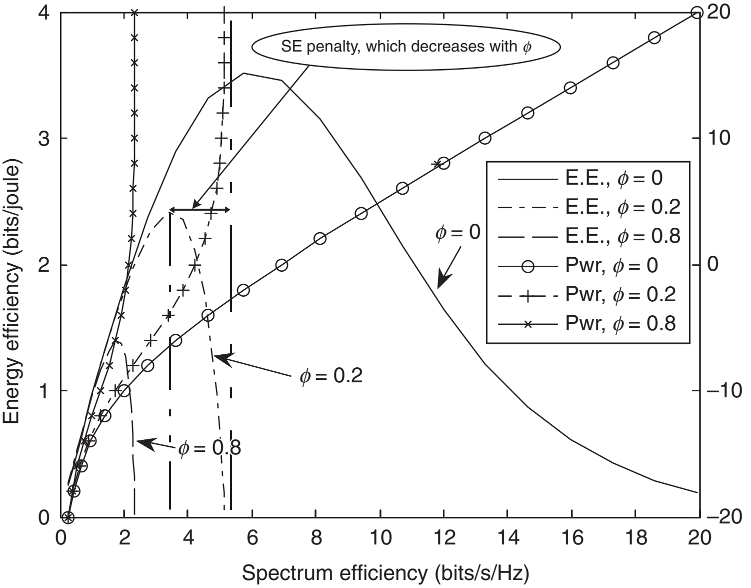


Figure 9.10 Tradeoff between energy and spectral efficiency (left *Y* axis: for curves without markers and indicates the achieved EE while the SE is as in *X* axis; right *Y* axis: for curves with markers and indicates the required transmission power to achieve the spectral efficiency given in *X* axis. $P_c = 1, g = 1, \sigma^2 = 0.01, N = 2$).

by choosing enough transmission power. When $\phi > 0$, the spectral efficiency beyond the spectral efficiency upper bound is not achievable. Energy efficiency is much more sensitive to the power selection than spectral efficiency. When $\phi = 0.1$, the transmission power should be -3 dBW for energy-efficient communications. The corresponding spectral efficiency is 4.2 bits/s/Hz and the energy efficiency is 2.8 bits/joule. If the transmission power is further increased, the energy efficiency decreases very fast while the spectral efficiency improves very slowly. Thus in interference-limited scenarios, increasing transmission power beyond what is needed for the highest energy efficiency improves spectral efficiency very little but will significantly hurt energy efficiency. In addition, the power optimization that achieves the highest energy efficiency also has reduced penalty in spectral efficiency with the increase of ϕ. We can see that energy-efficient communications have significant advantages, especially in heavy interference environments. For example, in cell-edge communications, energy-efficient power control has the potential to reduce device energy consumption without losing throughput.

In this section, we have made some simple assumptions to gain intuitive understanding about the tradeoff between energy and spectral efficiency. In practice, different users will experience different levels of interference from each other. In addition, some users may experience interference because they share the same channels, and others experience no interference as they use orthogonal channels. The tradeoff between energy and spectral efficiency in general will be the statistical average of the effects discussed in these two sections.

Table 9.1 Power consumption of a wireless transceiver.

	802.11b	802.11a	802.11g
Sleep mode	132 mW	132 mW	132 mW
Idle mode	544 mW	990 mW	990 mW
Receive mode	726 mW	1320 mW	1320 mW
Transmit mode	1089 mW	1815 mW	1980 mW
Data rate	11 Mbps	54 Mbps	54 Mbps

9.5 Energy-efficient MAC design

Each wireless device may have five operating modes: transmit, receive, idle, sleep and off. The main functions of each mode are listed below.

(i) Transmit mode: send data.
(ii) Receive mode: receive data.
(iii) Idle: all transceiver components are on and ready to send or receive data.
(iv) Sleep: the major transceiver circuit components are turned off with a very limited portion remaining on to listen to demands outside.
(v) Off: power completely off.

Of the five modes, transmit, receive and idle modes can be called active modes because all circuit components are running and these components will consume circuit power. The difference in the amount of energy consumed in these modes is significant. An example is given in Table 9.1, which lists the power consumption of several commercial 802.11 transceivers in all operation modes [R. Mangharam et al., 2005]. We can see that while the sleep mode power consumption stays the same, the power consumption in other modes has increased with each new standard that supports a higher data rate. The power consumption in transmit mode will be even higher for longer-distance communications, such as in cellular networks, as more power is needed to compensate for path loss, which grows exponentially with communication distances.

There are many sources of energy consumption from a MAC perspective. Below we list some examples.

- Traffic usually arrives sporadically but the transceiver circuit components need to be kept on even when there is no data so that the transceiver can respond to traffic quickly. Traffic examples include email, web page browsing, VoIP and so on. The device usually stays in idle mode for a long time running all circuits without any transceiving activities. The energy consumption in idle mode can be much larger than in other modes.
- In a wireless cellular network, the receiver has to be powered on all the time waiting to receive data. In the downlink of cellular networks, mobile devices have to stay on so that they can receive system messages from the base station. In addition it may have to decode packets destined to other users and therefore waste energy. In the uplink,

the base station also needs to be turned on even if there are no users being served so that future connection requests from mobile users can always be heard and the states of mobile users can be monitored.

- In wireless networks using random access protocols, collisions may occur and the data sent out will no longer be useful. The energy in transmission and reception will be wasted. Therefore it is necessary to avoid collisions or failures in packet reception as much as possible.
- When mobile devices switch between different modes, it takes some time and consumes a significant amount of transition energy. In addition to the turnaround time within the devices, the network also needs to be notified of the changes, which will take a much longer time.

Observing the above analysis, the main purpose of energy-efficient distributed MAC design is to enable mobile devices to stay in the least-power-consuming mode, for example sleep mode, as long as possible. With this design, the devices can deactivate as many functional circuit units as possible. However, there are issues related to switching off and on circuit components. For example, restarting will cause delay in responding to traffic requests. It may incur even higher energy consumption because of higher restarting current requirement. Thus there is a tradeoff between performance and energy saving.

If there is no activity for a certain amount of time, a device can enter sleep mode. A timer can therefore be used to determine when a component can be turned off. This timer can be either statically or dynamically configured, depending on the scheduling algorithm, and will affect system energy consumption. A device needs to wake up to respond to user or network activities from time to time. If the device is completely off, it will not be able to hear any activity request from outside and will never wake up again. A small portion of circuits therefore need to remain on to monitor incoming requests. It takes non-negligible time to wake up and the latency can be too long. Therefore it is necessary to predict when responses are needed and wake up in advance.

Communications with large packet sizes and stable traffic flows can enter sleep mode when there is no traffic for a while to save energy. Since less mode switching is needed, little performance loss is expected in mode switching. For communications with sporadic and frequent traffic arrivals, transitions between modes can be too frequent and sleeping whenever there is no data in the buffer will hurt the performance while not saving any energy. One improvement is to buffer the traffic and schedule transmission such that it can continuously transmit or receive data and sleep longer, thereby staying in one mode as long as possible. For example the data for different applications can be buffered, bundled, and delivered together to reduce the number of transitions.

To minimize the active time of all devices while ensuring the success of packet transmission, a distributed energy-efficient MAC protocol should command in advance when each device should send or receive data so that the device can choose to sleep in the remaining time. There are many ways of designing energy-efficient MACs. For example a base station can buffer the data and periodically broadcast a message indicating which mobile devices need to receive the buffered data. Each mobile device has to wake up to receive the broadcast message. If there is data for a device, it will wake

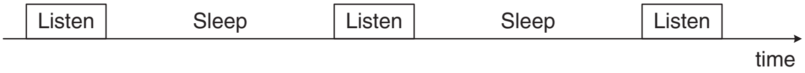


Figure 9.11 Periodic listen and sleep.

up and receive the data; otherwise it will sleep again. The synchronization between the base station and mobile devices enables mobile devices to sleep and wake up just in time for communications.

Distributed energy-efficient MAC design is widely used in wireless sensor networks. In these networks, each sensor node may consist of one or more sensors, an embedded processing unit and a low-power radio. Applications of sensor networks include earthquake monitoring in desert areas, traffic control, industrial automation, and so on. Most sensor nodes are usually battery powered and energy efficiency is thus of paramount importance because replacing batteries for all nodes frequently is difficult or even impossible. To save energy, sensor nodes will keep silent for most of the time and become active for data transmission after detecting something of interest. A typical example is sensor-MAC (S-MAC) [W. Ye et al., 2002]. To save power, S-MAC implements a periodic listen-and-sleep protocol, which is illustrated in Figure 9.11. Each node communicates with other nodes in the listening period and sleeps in the sleeping period. When the node sleeps, it will turn off its radio to save power. With the periodic design, a large amount of energy consumption can be saved by avoiding unnecessary idle listening, especially when traffic load is low. To make the protocol more flexible, a parameter, duty cycle, is defined as the ratio of the listen period to a complete sleep and listen cycle. The duty cycle can be adjusted from 1% to 100%. The listen period is divided into two parts, SYNC and DATA periods. The SYNC period is used to solve synchronization problems between neighboring nodes and the DATA period is used for data transmission. To ensure the success of packet transmission, S-MAC uses a CSMA/CA protocol with RTS/CTS to avoid collisions.

Example 9.3: A slotted ALOHA system has n terminals. At the beginning of each time slot, each terminal sends a packet of D bits to the common receiver with probability p, which includes both new packets and retransmitted packets. The time slots have length T. The transmission data rate is r. After each transmission attempt, the terminal is switched to and stays in reception mode to receive possible positive acknowledgements, which will be received within the same time slot if there are any. In the transmission mode, the transmission power is P and the circuit power is P_c. The power consumption in the reception mode is P_l. If a terminal has no packet to send in a time slot, it consumes sleeping power P_s. Determine the energy efficiency of the terminals.

Solution: The probability π for a packet to be successfully received is:

$$\pi = (1-p)^{n-1}.$$

The average number of bits successfully delivered in each time slot is:

$$D_s = D\pi = D(1-p)^{n-1}.$$

Per time slot, the average time T_{tx} for the terminal to be active for packet transmission is:

$$T_{tx} = p\frac{D}{r}.$$

The average time T_{rx} in the reception mode is:

$$T_{rx} = p(T - T_{tx}).$$

The average time in the sleeping mode is:

$$T_s = (1-p)T.$$

The energy efficiency u_e is thus determined by:

$$\begin{aligned} u_e &= \frac{D_s}{(P+P_c)T_{tx} + P_l T_{rx} + P_s T_s} \\ &= \frac{D(1-p)^{n-1}}{(P+P_c)p\frac{D}{r} + P_l p(T - p\frac{D}{r}) + P_s(1-p)T}. \end{aligned} \tag{9.44}$$

Note that the techniques introduced in Section 9.3 can be used in this system to enhance energy efficiency by choosing the transmission power adaptively.

Example 9.4: Consider the slotted ALOHA model in Section 3.4.1. Assume all the users are uniformly distributed over a circular cell area accessing a base station in the center. The cell area has a radius D_o. The users use a power control scheme such that a constant signal power, P_r, is received at the base station. The signal level decays as the αth power of the distance, i.e. $P_r = \frac{P_T}{d^\alpha}$, where P_T is the transmitted power. If a user sends a packet in a time slot, the full time slot will be used for data transmission. The power amplifier efficiency of each user is ξ. When sending data in the active mode, each user consumes P_c circuit power. Each user goes to the sleep mode immediately when there is no data transmission and consumes only P_S circuit power. Similarly, a sleeping user can switch to the active mode and send data immediately when it needs to send a packet.

(i) Determine the cumulative distribution function (CDF) of the transmission power of a user when the user is sending data.

(ii) If there is only one user in the network, determine the CDF of the individual average power consumption.

(iii) If there are N users in the network, determine the CDF of the individual average power consumption of all users.

Solution: The CDF of the transmission power of all users is

$$Pr(P_T \leq P) = Pr(d^\alpha \leq \frac{P}{P_r}) = \frac{\sqrt[\alpha/2]{P}}{D_o^2 \sqrt[\alpha/2]{P_r}}. \tag{9.45}$$

Each device needs to be active only in the slots when there are packets to be sent. If there is only one user in the network, the average power consumption is

$$\overline{P} = \lambda\left(\frac{P_T}{\xi} + P_c\right) + (1-\lambda)P_S, \tag{9.46}$$

and the CDF of the average power consumption is

$$\begin{aligned} Pr(\overline{P} \le P) &= Pr\left(\lambda\left(\frac{P_T}{\xi} + P_c\right) + (1-\lambda)P_S \le P\right) \\ &= \frac{\sqrt[\alpha/2]{\left(\frac{P-(1-\lambda)P_S}{\lambda} - P_c\right)\xi}}{D_o^2\ \sqrt[\alpha/2]{P_r}}. \end{aligned} \tag{9.47}$$

When there are N users in the network, the active time slots of each user consist of those for both new transmissions and retransmissions. Therefore γ determines the probability of a user being in the active mode or not in each time slot. γ is determined by (3.30). As illustrated in Figure 3.17, for a certain traffic arrival rate $\lambda = \lambda_0$, there might be two γ values; one is comparatively low and smaller than 1, whereas the other is considerably higher. These two values correspond to states of equilibrium in the network. The state corresponding to the smaller γ value has fewer transmission activities and almost all transmissions are successful. In the other state with much larger γ, there are a large number of backlogged packets, causing frequent collisions. The attempt rate γ is very high with a very low success probability q. If the system is in the equilibrium state with lower γ value, the CDF of the average power consumption is

$$\begin{aligned} Pr(\overline{P} \le P) &= Pr\left(\gamma\left(\frac{P_T}{\xi} + P_c\right) + (1-\gamma)P_S \le P\right) \\ &= \frac{\sqrt[\alpha/2]{\left(\frac{P-(1-\gamma)P_S}{\gamma} - P_c\right)\xi}}{D_o^2\ \sqrt[\alpha/2]{P_r}}. \end{aligned} \tag{9.48}$$

Otherwise, $\gamma > 1$, indicating all users are always in the transmission state and the CDF of the average power consumption is

$$\begin{aligned} Pr(\overline{P} \le P) &= Pr\left(\left(\frac{P_T}{\xi} + P_c\right) \le P\right) \\ &= \frac{\sqrt[\alpha/2]{(P-P_c)\xi}}{D_o^2\ \sqrt[\alpha/2]{P_r}}. \end{aligned} \tag{9.49}$$

Because of the stochastic variations in the traffic arrival, the system may move between the two equilibrium states. The long-term average power consumption will thus depend on how often the two equilibrium states exist.

Figure 9.12 illustrates the CDFs of average power consumption when there is either one or the maximum number of users in the network. In the second case, $\gamma = 1$. As shown in the figure, almost ten times more power is consumed when there is the maximum number of users in the network.

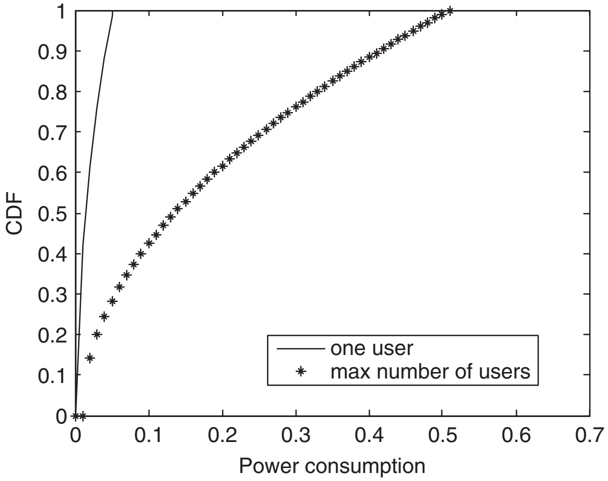


Figure 9.12 CDF of average power consumption ($\alpha = 4, P_c = 4$ mW, $P_S = 1$ mW, $\xi = 0.5, R = 100$ m, $\lambda_i = 0.025$).

9.6 Energy-efficient network management

In this section we discuss energy-efficient designs for wireless cellular networks, which is a major concern for operators to reduce operational expenditure and environmental impacts.

9.6.1 Energy-efficient network deployment

In this section we discuss how to deploy cellular networks in an energy-efficient way. Only path losses will be considered for simplicity.

In the downlink of a cellular network that is coverage limited, the received power at a distance d from the base station is

$$P_r = \frac{P}{d^\alpha}, \tag{9.50}$$

where P is the transmission power and α the path loss exponent. The interference is negligible and the SNR is

$$\Gamma = \frac{P}{d^\alpha N_0 W}, \tag{9.51}$$

where N_0 is the noise spectral density. The interference from adjacent cells is neglected to simplify the analysis. The data rate is given by

$$r = W \log_2 \left(1 + \frac{\Gamma}{\theta}\right). \tag{9.52}$$

Suppose cell-edge users, with $d = D_o$, desire a minimum data rate r_0. Solving the above equation for the transmission power, we have the transmission power

$$P = \left(2^{\frac{r_0}{W}} - 1\right)\theta D_o^{\alpha} N_0 W. \tag{9.53}$$

From (9.53), the required transmission power for a certain coverage increases exponentially with the cell coverage D_o. Now let's take a look at the total network energy consumption in a certain area. The number of base stations in a service area is given by

$$N_{BS} = \frac{A}{\pi D_o^2}, \tag{9.54}$$

where A is the area size.

The total energy consumption by the network for a duration t is

$$E_{net} = P_{net} t, \tag{9.55}$$

where P_{net} is overall network power consumption consisting of both communication and computation.

In the following we consider two cases for minimizing the energy consumption: power minimization and energy efficiency maximization.

(i) Power minimization: If the network keeps transmitting signals for a fixed time duration t and does not care if the data of the users has been delivered or not, minimizing energy consumption is the same as minimizing power consumption.

(ii) Energy efficiency maximization: If the network is designed to deliver the data of all users and then shut down the transmission modules to switch to a low-power mode, e.g. sleep mode, minimizing energy consumption is equivalent to maximizing network energy efficiency. Assume all the cell-edge users have Q bits to send and the time needed to send these bits is

$$t = \frac{Q}{N_{BS} r_0}; \tag{9.56}$$

the total energy consumption will be

$$E_{total} = P_{total} \frac{Q}{N_{BS} r_0} = \frac{Q}{u_e}, \tag{9.57}$$

where u_e is the network energy efficiency defined as

$$u_e = \frac{N_{BS} r_0}{P_{total}}. \tag{9.58}$$

So minimizing energy consumption is equivalent to maximizing network energy efficiency, which is the same as the link design in Section 9.3.

Communications energy consumption

The total required transmission power in a cellular network to meet the coverage requirement can be written as a function of N_{BS} as

$$P_{total} = N_{BS}P = \left(2^{\frac{r_0}{W}} - 1\right)\theta N_0 W \left(\frac{A}{\pi}\right)^{\frac{\alpha}{2}} N_{BS}^{1-\frac{\alpha}{2}}. \tag{9.59}$$

In the power minimization case, we can see that the total transmission power in (9.59) of a mobile radio network with a fixed service area decreases with the number of base stations in the network for any $\alpha > 2$. In practice, α is always bigger than 2. Therefore if we consider only transmission power, the cell size should be as small as possible. Hence, a high-density deployment strategy with many micro base stations consumes less energy.

In the case for maximizing energy efficiency,

$$u_e = \frac{N_{BS}r_0}{P_{total}} = \frac{r_0 N_{BS}^{\frac{\alpha}{2}}}{\left(2^{\frac{r_0}{W}} - 1\right)\theta N_0 W \left(\frac{A}{\pi}\right)^{\frac{\alpha}{2}}}. \tag{9.60}$$

The energy efficiency of the mobile radio network with a fixed service area increases with the number of base stations. If we consider only transmission power, the cell sizes should be as small as possible. Hence, a high-density deployment strategy with many micro base stations is also the most energy efficient.

The user requirement also has an impact on network energy consumption. Both P_{total} and u_e are strictly decreasing in r_0. If users only desire best-effort service and r_0 can be any value, the network energy efficiency is maximized when r_0 is as small as possible. If users have a minimum data requirement $\bar{r}_0$, then the power should be selected so that the cell edge rate is $\bar{r}_0$.

Total energy consumption

In practice, in addition to communications energy consumption, each base station also consumes computation energy for signal processing, site cooling, backhaul and so on. Then the total network power consumption can be modeled by

$$\begin{aligned} P_{net} &= \frac{P_{total}}{\zeta} + N_{BS}P_c + P_o \\ &= \left(2^{\frac{r_0}{W}} - 1\right)\theta N_0 W \left(\frac{A}{\pi}\right)^{\frac{\alpha}{2}} N_{BS}^{1-\frac{\alpha}{2}}/\zeta + N_{BS}P_c + P_o, \end{aligned} \tag{9.61}$$

where P_o is the background power consumption in the network independent of the number of base stations.

In the power minimization case, it can be easily proved that (9.61) is strictly convex. There exists a unique finite optimal N_{BS} that minimizes (9.61), as shown in Figure 9.13. The optimal N_{BS} and the corresponding minimum network power consumption can be

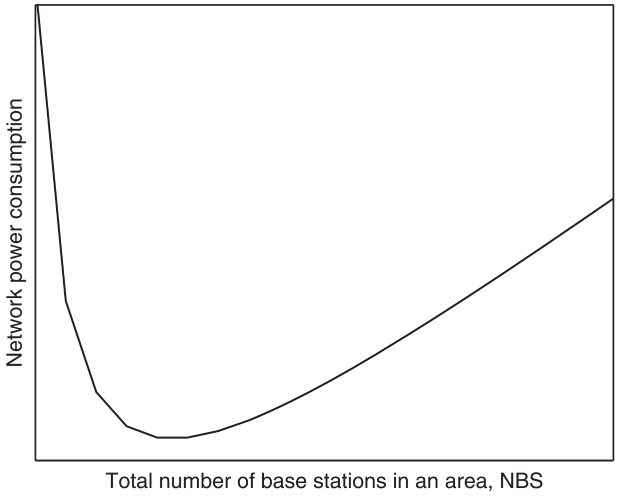


Figure 9.13 Network power consumption and base station density relationship with a cell edge rate requirement.

found by setting the first-order derivative to be zero and

$$\frac{N_{BS}^*}{A} = \frac{1}{\pi}\left(\frac{P_\beta}{P_c}\right)^{\frac{2}{\alpha}}\left(\frac{\alpha}{2}-1\right)^{\frac{2}{\alpha}}, \tag{9.62}$$

where $P_\beta = \left(2^{\frac{r_0}{W}}-1\right)\theta N_0 W/\zeta$, the desired receiving power of a cell-edge user if the power amplifier efficiency is one.

However, in the second case, the network energy efficiency is

$$u_e = \frac{N_{BS}r_0}{P_{net}} = \frac{r_0}{\left(2^{\frac{r_0}{W}}-1\right)\theta N_0 W\left(\frac{A}{\pi}\right)^{\frac{\alpha}{2}} N_{BS}^{-\frac{\alpha}{2}}/\zeta + P_c + P_o N_{BS}^{-1}}, \tag{9.63}$$

which strictly increases in N_{BS}. Hence a high-density deployment strategy is always desired for energy-efficient communications. While more base stations will increase the network power consumption, the additional capacity generated by using the additional base stations will offset the energy cost and further increase the network energy efficiency. The network energy efficiency is bounded above by

$$u_e^* = \lim_{N_{BS}\to\infty} u_e = \frac{r_0}{P_c}, \tag{9.64}$$

which is achieved when the number of base stations is sufficiently large and the transmission energy consumption is close to zero.

While P_{net} is strictly decreasing in r_0, energy efficiency u_e is not. u_e is strictly quasi-concave in r_0 and the optimal r_0^* that maximizes u_e can be found using Theorem 9.2. If $r_0^* \leq r_0$, the transmission power of the base stations should be the one in (9.53). Otherwise, the transmission power should be

$$P^* = \left(2^{\frac{r_0^*}{W}}-1\right)\theta D_o^\alpha N_0 W, \tag{9.65}$$

which enables a higher data rate so that the base stations can finish the data transmission sooner to save energy.

Note that the high-density deployment strategy is only a theoretical one to gain insights into the problem. When the network gets more and more crowded, the network will be more and more interference-limited and a different approach should be used to analyze the optimal deployment strategy. In addition, the number of users in an area is also limited and it would make no sense when the base station density goes beyond a certain number.

9.6.2 Heterogeneous network deployment

A macro cell usually covers a large area and is therefore not efficient in providing broadband services. According to the discussion in Section 9.6.1, one way to improve energy efficiency is to decrease the coverage of cells and thus reduce the signal propagation loss to reduce transmission power. Heterogeneous cellular deployment with small cells, for example micro-, pico- or femtocells, under umbrella macro cells can therefore be used to improve network energy efficiency. An example is illustrated in Figure 9.14. The small cells are cells served with lower-power base stations. A micro-or pico-cell usually covers a range of one or several hundred meters and can be used to cover smaller hotspot areas with dense traffic, such as shopping malls, hotels, airports, train stations and so on. A femtocell can be used to cover a much smaller area like an individual house and the coverage can be only a few meters or ten meters. Small cells are much more power efficient than macro cells. For example, a typical femtocell may consume only 5 W in total compared to several thousand watts that would be needed

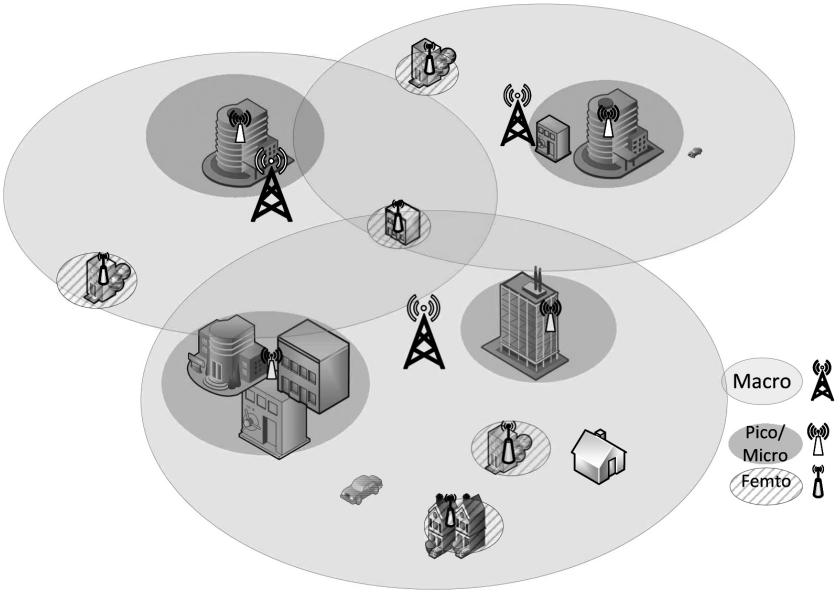


Figure 9.14 An example of heterogeneous networks.

to support a macro cell. The optimal deployment of small cells may follow a similar strategy to those discussed in Section 9.6.1.

9.6.3 Energy-efficient cellular network operation

Traffic load in cellular networks varies significantly in time and space. For example, traffic is usually much heavier in office areas than residential areas in the daytime of weekdays and the other way around after work. Therefore many cells may be heavily loaded at one time and carry almost no traffic at another, and vice versa. The traffic fluctuations would be much higher for small cells.

Traditional access networks are planned based on the peak hour traffic, and static cell sizes are usually used. With static cell size deployment, the network does not adapt to the fluctuating traffic loads and always works with peak power consumption to achieve the highest performance. To save energy, cell size can therefore be adjusted depending on traffic in the network. This technique is used in CDMA networks, in which cells with heavy loads will reduce their cell sizes by decreasing the transmission power through power control. Some users at cell edges will be handed off to adjacent cells with lower traffic and the traffic load is therefore shared across adjacent cells. Besides, handing over cell-edge users from heavy-traffic cells to light-traffic ones also saves transmission power. This is because these users will be allocated more spectrum resources in the light-traffic cells and therefore energy efficiency can be improved, according to Proposition 9.3. The technique is frequently called cell-breathing. To achieve the most energy saving, network-level power management will be needed so that multiple cells can coordinate to decide traffic handovers.

Running a base station consumes a considerable amount of energy and choosing some base stations to sleep if their traffic loads are very low can save a significant amount of energy. When some base stations are in sleep mode, their coverage can be preserved by the remaining adjacent active cells. There are several techniques to enlarge the coverage of the adjacent active cells such that they cover the cells in sleep mode. For example, increasing the transmission power or adjusting the antenna tilt can expand their coverage, as shown in Figure 9.15. Furthermore, several adjacent base stations can cooperate to cover the area of cells in sleep mode using the so-called Coordinated Multi-Point (CoMP) transmit/receive technology. With CoMP, the reception power can

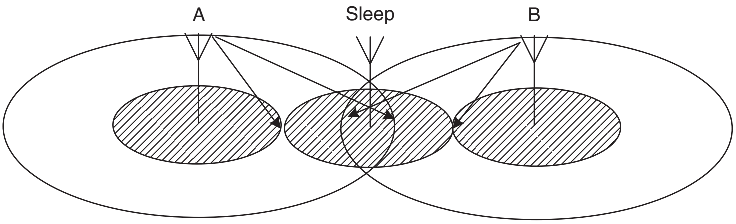


Figure 9.15 Coverage preserved by adjacent cells.

be increased because signals from several cells can be combined together for signal detection, resulting in an expansion of cell coverage.

Exercises

9.1 Prove that both the throughput and energy efficiency of a wireless transmitter increase with the signal bandwidth in an AWGN channel.

9.2 A wireless terminal consumes 2 W circuit power in transmission mode. In sleep mode, it consumes 0.5 W circuit power. It needs to send 100 bits in the buffer over an AWGN channel once every T seconds. After sending all bits, it goes to sleep mode to save power. Assume the transitions between different modes are instantaneous. All other parameters, if not specified, are assumed to have a value of 1. How fast, in terms of both data rate and time, should the terminal send the bits?

9.3 A slotted ALOHA system has five terminals. At the beginning of each time slot, each terminal sends a packet of 10 bits to the common receiver with probability 0.15, which includes both new packets and retransmitted packets. The time slots have length T. After each transmission attempt, the terminal is switched to and stays in reception mode to receive possible positive acknowledgements, which will be received within the same time slot if there are any. In transmission mode, the circuit power is 100 mW. The transmission power P_t is determined by the data rate, r, of the terminal when sending the packet and is given by $r = \log_2(1 + P_t)$. The power consumption in reception mode is 100 mW. If a terminal has no packet to send in a time slot, it consumes sleeping power 20 mW. Determine the data rate that the terminals should use to minimize their energy consumption and the corresponding energy efficiency of the terminals when a) $T = 10$ ms, or b) $T = 10$ s.

9.4 In a slotted ALOHA system collided packets are all lost and single transmitted packets are assumed to be lost with probability 20% due to noise. Each transmitter needs to send a 15 kbit message on average every 10 seconds following the Poisson process, even if the previous messages have not yet been sent. Assume the compound process of retransmitted and new messages will also form a Poisson process. Assume the transmitters know if a packet is successfully received or not immediately after its transmission. When sending a packet, the transmission data rate of the transmitter is determined by $R = 10 \times SNR$ kbps. The transmission power P_t is determined by $P_t = P_r \times d^4/z$, where P_r is the constant received power and $P_r = 15e^{-4}$ mW. $z = 0.2$ is the power amplifier efficiency. The noise power is $No = 1e^{-4}$ mW. d is the distance between the transmitter and receiver of each communicating pair and is uniformly distributed between 10 and 100 meters. Each transmitter consumes 200 mW circuit power in the data transmission mode. If a transmitter is not transmitting data, it always sleeps and consumes 20 mW circuit power in the sleep mode. There is no delay or additional energy consumption in switching between the different modes.

(a) Determine the CDF of transmission power of the transmitters in the transmission mode.
(b) Determine the CDF of average power consumption of a transmitter if there is only one communicating pair.
(c) If there is only one communicating pair with $d = 50$ meters, what is the energy efficiency of the transmitter?
(d) How many communicating pairs can this system support?
(e) When the maximum number of pairs given in (d) exist in the network, determine the energy efficiency of a transmitter which is 50 meters distant from its receiver.

References

M. S. Bazaraa, H. D. Sherali and C. M. Shetty. 2013. *Nonlinear Programming: Theory and Algorithms*. Chichester: Wiley.

W. Dinkelbach. 1967. On nonlinear fractional programming. *Management Science*, 13(7), 492–498.

EE. Times 2007 (Sept.). Green issues challenge base station power. *EE Times*.

GeSI. 2008. *Smart 2020: Enabling the low carbon economy in the information age*. London: The Climate Group.

C. Isheden and G. P. Fettweis. 2011 (Mar.). Energy-efficient link adaptation with transmitter CSI. Pages 1381–1386 of: *Proc. IEEE WCNC 2011*.

K. Lahiri, A. Raghunathan, S. Dey and D. Panigrahi. 2002 (Jan.). Battery-driven system design: A new frontier in low power design. Pages 261–267 of: *Proc. Intl. Conf. on VLSI Design*.

R. Mangharam, R. Rajkumar, S. Pollin, F. Catthoor, B. Bougard, L. Van der Perre and I. Moeman. 2005 (Mar.). Optimal fixed and scalable energy management for wireless networks. Pages 114–125 of: *Proc. IEEE INFOCOM 2005*, vol. 1.

G. W. Miao, N. Himayat, Y. Li and D. Bormann. 2008 (May). Energy-efficient design in wireless OFDMA. Pages 3307–3312 of: *Proc. IEEE ICC 2008*.

G. W. Miao, N. Himayat and Y. Li. 2010. Energy-efficient link adaptation in frequency-selective channels. *IEEE Trans. Commun.*, 58(2), 545–554.

G. W. Miao, N. Himayat, Y. Li and S. Talwar. 2011. Distributed interference-aware energy-efficient power optimization. *IEEE Trans. Wireless Commun.*, 10(4), 1323–1333.

G. W. Miao, N. Himayat, G. Y. Li and S. Talwar. 2012. Low-complexity energy-efficient scheduling for uplink OFDMA. *IEEE Trans. Commun.*, 60(Jan.), 112–120.

G. W. Miao and G. Song. 2015. *Energy and Spectrum Efficient Wireless Network Design*. Cambridge: Cambridge University Press.

W. Ye, J. Heidemann and D. Estrin. 2002 (June). An energy-efficient MAC protocol for wireless sensor networks. Pages 1567–1576 of: *Proc. IEEE Infocom 2002*.

10 Long term evolution

3GPP Long Term Evolution (LTE) represents the fourth generation of cellular technologies. It is designed to support high-speed multimedia unicast and broadcast services. The LTE physical layer is very efficient in handling both data and control signaling and employs advanced technologies like orthogonal frequency division multiplexing (OFDM) and multiple input multiple output (MIMO). In the downlink, LTE uses orthogonal frequency division multiple access (OFDMA) and in the uplink, single carrier-frequency division multiple access (SC-FDMA). LTE allows flexible resource allocation on a subcarrier-by-subcarrier basis for a specified number of OFDM symbols, which significantly increases spectral efficiency. In addition, LTE implements advanced interference management schemes to boost overall network capacity. Both frequency division duplexing (FDD) and time division duplexing (TDD) are supported in LTE. This chapter will focus on LTE FDD systems and the corresponding radio resource management in the physical layer. The goal of this chapter is not to exhaust all tutorial information on LTE, but rather to illustrate the combination of underlying theoretical principles introduced in the previous chapters of this book and the specific system design constraints in LTE.

10.1 Physical layer for downlink

The LTE downlink transmission multiplexes both UE (user equipment) data and control signaling. There are three dimensions in the downlink transmission resources: time, frequency and space. The time–frequency resources are divided using orthogonal frequency division multiple access (OFDMA) and the resources in the spatial dimension are managed by multiple antenna transmission and reception techniques. Below we will give a brief introduction to the key technologies in forming the transmitted downlink signals and the resource structure in LTE.

10.1.1 Orthogonal frequency division multiplexing

The main advantage of using OFDM is to increase robustness against frequency-selective fading and narrowband interference. OFDM is a modulation scheme that fits high-speed communications in delay-dispersive environments as it converts a high-data-rate stream into many low-rate streams. These low-rate streams are transmitted over parallel, orthogonal, narrowband channels that can be easily equalized. OFDM was first invented in the mid-1960s [R. W. Chang, 1966, 1970]. In 1985, Cimini was the first to describe the use of OFDM for wireless communications [L. J. Cimini, 1985]. In this section we give a brief introduction to the basic properties and advantages of OFDM.

As shown in Figure 3.6, the bit stream is first converted to K parallel streams using a serial to parallel (S/P) converter. The K streams are independently coded and modulated, resulting in K streams of complex data symbols. Each stream will be transmitted independently on the corresponding OFDM subcarrier. Different streams may use different modulations, for example QPSK or 16QAM, and coding rates. Because of the frequency selectivity, the channel gains of different subcarriers may be different and thus some streams may achieve higher data rates than others.

Let $\mathbf{b}[n] = [b_0[n], b_1[n], \ldots, b_{K-1}[n]]^T$ denote the vector of complex symbols for the nth OFDM symbol, where $b_i[n]$ is the complex symbol of the ith stream. $\mathbf{b}[n]$ is then processed with an N-point inverse FFT (IFFT), resulting in N complex time-domain signals $\mathbf{d}[n] = [d_0[n], d_1[n], \ldots, d_{N-1}[n]]^T$. Usually $N \geq K$ and the unmodulated subcarriers are padded with zeros, for example

$$\tilde{\mathbf{b}}[n] = [b_0[n], b_1[n], \ldots, b_{K-1}[n], 0, \ldots 0, 0]^T \tag{10.1}$$

is used to obtain $\mathbf{d}[n]$.

Before sending $\mathbf{d}[n]$ out, a guard period is created at the beginning of $\mathbf{d}[n]$ to eliminate the inter-symbol interference caused by the multipath propagation effect. The guard period is obtained by adding the cyclic prefix (CP) to the beginning of $\mathbf{d}[n]$. Assume the guard period has G samples. The CP is the duplicate of the last G elements of $\mathbf{d}[n]$. Therefore, as shown in Figure 10.1, the complete time-domain OFDM symbol including the CP is

$$\tilde{\mathbf{d}}[n] = [d_{N-G}[n], \ldots, d_{N-1}[n], d_0[n], d_1[n], \ldots, d_{N-1}[n]]^T. \tag{10.2}$$

$\tilde{\mathbf{d}}[n]$ is then converted from parallel to serial for RF transmission.

Assume the receiver is synchronized in both the time and frequency domains using reference signals. The reverse operations can be used by the receiver to decode the OFDM symbol. The first G CP samples of the received signal are removed and the remaining N samples are passed to the FFT for the transformation back to the frequency domain. The symbols on the modulated K subcarriers are chosen out of the N output

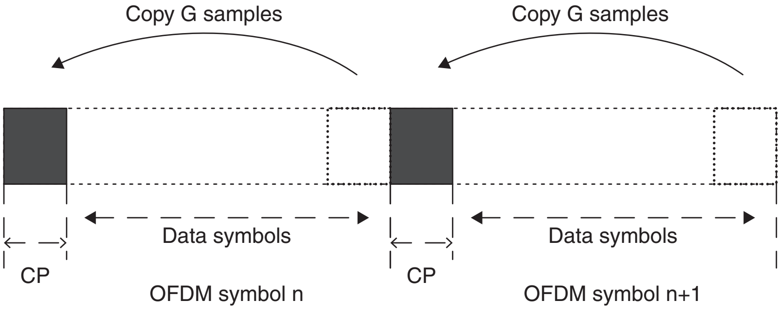


Figure 10.1 Cyclic prefix insertion.

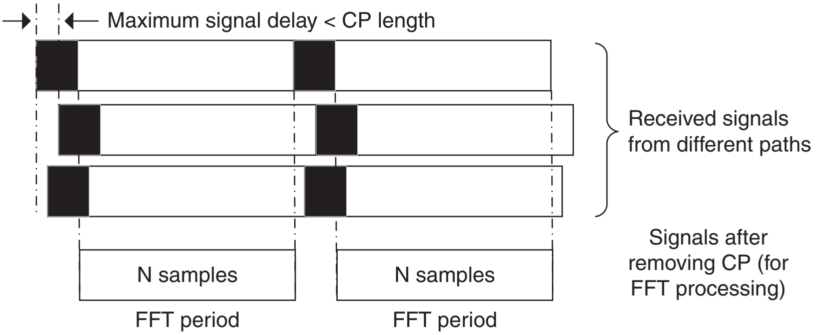

Maximum signal delay < CP length
Received signals from different paths
Signals after removing CP (for FFT processing)
N samples
N samples
FFT period
FFT period

Figure 10.2 Inter-symbol interference elimination via CPs.

subcarriers and are further processed, for example demodulated and decoded, for the desired bit stream.

Because of the multipath effect, the receiver may receive several replicas of the transmitted signals at different delays, as shown in Figure 10.2. Therefore it is possible to have interference from the preceding symbol. The CP should be longer than the longest channel impulse response so that none of the replicas of the preceding symbol may spill over into the FFT period of the current symbol. After discarding the CP samples, there are only signals of time-staggered replicas of the current symbol, based on which the receiver can apply an equalizer to recover the original transmitted signal. Further analysis will show that the CP of OFDM changes the linear convolution of the multipath channel into a circular one [S. Stefania et al., 2009]. By the FFT operation, the circular convolution is transformed into a multiplicative operation in the frequency domain. Hence the frequency-selective multipath channel is converted into N parallel orthogonal flat-fading channels in the frequency domain. That is,

$$y_k[n] = b_k[n]H_k[n] + N_k[n], \tag{10.3}$$

where $H_k[n]$ is the channel gain on the kth subcarrier, $N_k[n]$ the corresponding noise and $y_k[n]$ the received signal to be equalized. The receiver will estimate the channel based on known reference signals that are periodically transmitted. The channel distortion can be corrected on a per-subcarrier basis, for example by applying an amplitude and phase shift on each subcarrier.

Example 10.1: Consider an OFDM system with 1024 active subcarriers with 1.5625 ms guard interval. The space between adjacent subcarriers is 160 Hz. 4QAM is used for the modulation on each subcarrier. A 2048-point IFFT is used to create the OFDM symbol.

(i) What are the system bandwidth and data rate?

(ii) What is the sampling frequency of the output symbol? How many samples are in the guard interval?

Solution: (i) The system bandwidth is

$$B = N \times \triangle f = 1024 \times 160 = 163{,}840\ Hz.$$

The symbol length is

$$T_s = \frac{1}{\triangle f} = \frac{1}{160} = 0.00625\ s.$$

The system data rate is

$$R = \frac{\log_2(M) \times N}{T_s + T_g} = \frac{2 \times 1024}{0.00625s + 0.0015625} = 18,774,681\ b/s.$$

(ii) The sampling frequency is

$$f_s = \frac{N_f}{T_s} = 2048/0.00625 = 327,680\ Hz.$$

The number of samples in the guard interval is

$$N_g = \frac{N_f T_g}{T_s} = 2048 \times 0.0015625/0.00625 = 512.$$

10.1.2 Orthogonal frequency division multiple access

The LTE downlink employs OFDMA as the multiplexing scheme because of its efficiency in managing resources and low latency in sending packets. In OFDM, a single item of user equipment (UE) communicates on all the subcarriers at any time. In OFDMA, UEs are allocated a specific number of subcarriers for a predetermined amount of time by the base station, called eNodeB in LTE, so that multiple UEs can be scheduled for data transmission simultaneously. The scheduling resource unit is referred to as a physical resource block (PRB) in LTE and, as illustrated in the following sections, a PRB has both time and frequency dimensions.

Figure 10.3 illustrates the physical layer frame structure with FDD. A different frame structure is defined for TDD, which is beyond the scope of this chapter. As shown in Figure 10.3, each frame has a duration of 10 ms and is divided into ten subframes. Each subframe is further divided into two slots of 0.5 ms length. Each slot consists of six or seven OFDM symbols, depending on how long the CP is. In Figure 10.3, a normal short CP is used.

The number of subcarriers depends on the system bandwidth. In LTE, the system bandwidth can be 1.4 MHz to 20 MHz and the number of available PRBs can be 6 to 100, as shown in Table 10.1. Each PRB consists of 12 consecutive subcarriers for one slot. Figure 10.4 illustrates the resource grid of the LTE downlink. Each box in the grid represents a single subcarrier for one symbol period and is called a resource element (RE). Therefore each PRB is composed of 84 REs. When spatial multiplexing is used, there is a resource grid for each spatial layer, which will be discussed in more detail in Section 10.1.3. A PRB is the smallest scheduling element of the base station. The scheduling is based on the channel state information of all UEs and the PRBs of different UEs may have different channel qualities. Adaptive allocation can therefore be performed to improve the total network spectral efficiency compared to single-user

Table 10.1 Available system bandwidth and RBs.

Channel bandwidth (MHz)	1.4	3	5	10	15	20
Number of PRBs	6	15	25	50	75	100

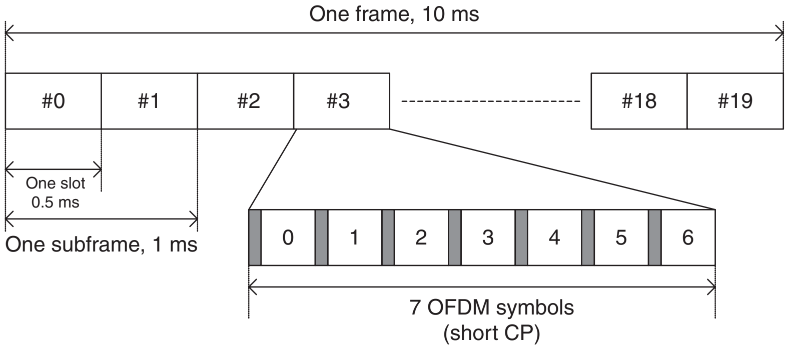


Figure 10.3 LTE generic frame structure.

OFDM systems by exploiting multiuser diversity. The minimum scheduling interval is 1 ms, corresponding to the 1 ms subframe, to enable low latency.

In LTE, there are no preambles for synchronization or channel estimation. Instead, the reference signals are embedded in the PRBs following a certain pattern and are used by the receiver for these purposes. An example of the reference signal pattern is shown in the grey REs in Figure 10.4. While the channel responses on REs carrying the reference signals can be estimated directly, the responses on other REs need to be estimated using interpolation.

In addition to reference signals, some REs in certain PRBs are reserved for synchronization signals, control signalling or cell broadcast information. The remaining REs are used for the transmission of the payload of LTE terminals, known as user equipments (UEs).

10.1.3 Multiple antenna techniques

Equipping wireless systems with multiple antennas is a powerful means of improving network performance. MIMO (multiple input, multiple output) is a multiple antenna technology in which multiple antennas are used at both the transmitter and the receiver. While recognized very early, the utilization of multiple antenna techniques in the mass commercial market started only around 2000. MIMO was adopted for the first time in a mobile network standard in Release 7 of HSDPA and LTE was the first global mobile system to be designed with MIMO as a key component from the start [S. Stefania et al., 2009].

Both the base station and UEs may be equipped with multiple antennas. If a base station is serving only one UE in a time–frequency resource, it is called single-user MIMO (SU-MIMO), otherwise, multi-user MIMO (MU-MIMO) as multiple UEs

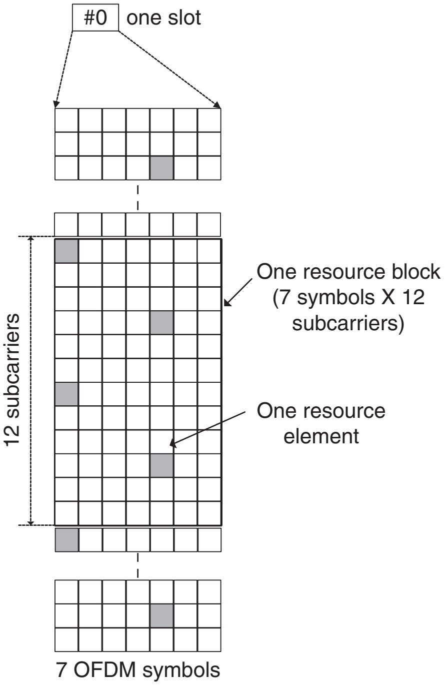


Figure 10.4 LTE downlink resource grid.

are being served simultaneously on the resource. The designs of SU-MIMO and MU-MIMO are based on some common fundamental principles of MIMO and mainly exploit the diversity gain or spatial multiplexing gain of multi-antenna systems to increase link robustness or data rates substantially.

LTE supports different MIMO modes. For example,

(i) single-antenna transmission
(ii) transmit diversity
(iii) open-loop spatial multiplexing (no CSI feedback needed)
(iv) closed-loop spatial multiplexing (CSI feedback needed)
(v) multi-user MIMO.

In the following, we give a brief introduction to the transmit diversity and spatial multiplexing modes. Readers interested in other modes may refer to [S. Stefania et al., 2009] for the details.

In the following we consider a generic MIMO system with m transmit antennas and n receive antennas, as shown in Figure 10.5. In case the receive antennas are located at different UEs, it is downlink multi-user MIMO as the system may multiplex the downlink transmission of multiple UEs on the same time–frequency resources by using proper MIMO signal processing.

Let $\mathbf{x} = [x_1, x_2, \ldots, x_m]^T$ denote the signals emitted from the transmit antennas and $\mathbf{h}$ the $n \times m$ channel matrix, in which h_{ji} is the independent channel gain from the

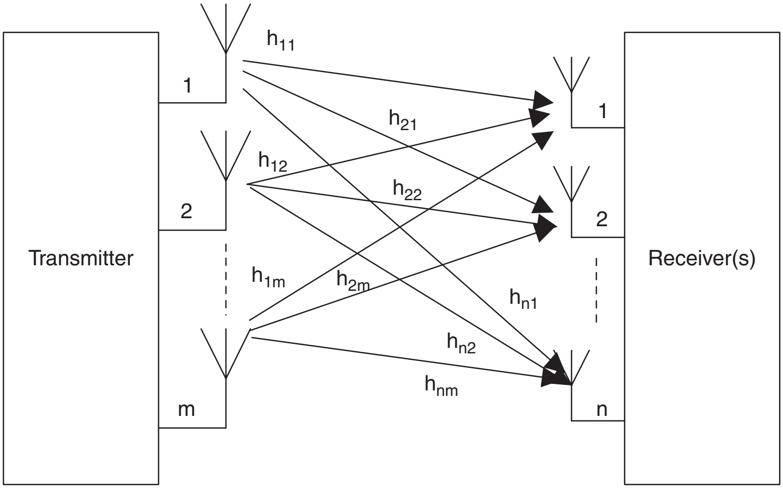


Figure 10.5 An $m \times n$ MIMO system.

ith transmit antenna to the jth receive antenna over a certain frequency subcarrier at a certain time slot. The subcarrier index and time slot index are omitted here for simplicity. The received signal is defined as **y** and

$$\mathbf{y} = \mathbf{hx} + \mathbf{N}, \tag{10.4}$$

where **N** is the zero-mean complex Gaussian noise on all antennas. Using the singular value decomposition (SVD) theorem, **h** can be decomposed as

$$\mathbf{h} = \mathbf{UDV}^H, \tag{10.5}$$

where **U** is an $n \times n$ unitary matrix and **V** an $m \times m$ unitary matrix. **D** is an $n \times m$ non-negative diagonal matrix, in which the diagonal entries are the square root of the eigenvalues of $\mathbf{hh}^H$. Here $(\mathbf{A})^H$ is the Hermitian of matrix **A**. Now we have the received signal in the following format:

$$\mathbf{y} = \mathbf{UDV}^H\mathbf{x} + \mathbf{N}. \tag{10.6}$$

Left multiply both sides by $\mathbf{U}^H$ and we have

$$\mathbf{U}^H\mathbf{y} = \mathbf{U}^H\mathbf{UDV}^H\mathbf{x} + \mathbf{U}^H\mathbf{N} = \mathbf{DV}^H\mathbf{x} + \mathbf{U}^H\mathbf{N}. \tag{10.7}$$

Define $\tilde{\mathbf{y}} = \mathbf{U}^H\mathbf{y}$, $\tilde{\mathbf{x}} = \mathbf{V}^H\mathbf{x}$ and $\tilde{\mathbf{N}} = \mathbf{U}^H\mathbf{N}$. The following equivalent system model can be obtained:

$$\tilde{\mathbf{y}} = \mathbf{D}\tilde{\mathbf{x}} + \tilde{\mathbf{N}}. \tag{10.8}$$

$\tilde{\mathbf{N}}$ is an $n \times 1$ zero-mean complex Gaussian noise vector. We can see that the MIMO transmission is decomposed into $r = \min(m,n)$ parallel independent transmissions, which is the basis for the spatial multiplexing, as will shortly be discussed.

Transmit diversity

In the transmit diversity mode, MIMO is used to exploit the spatial diversity and increase the transmission robustness. In this mode, signals are transmitted from multiple antennas simultaneously. Each antenna transmits the same information, so replicas of the same signal are received, which improves the receiver SNR and the robustness of data transmission. This is because different antennas experience independent channel fading and the probability that all signal replicas are in deep fades at the same time is very low. So the performance can be improved significantly by combining the received signal replicas using intelligent algorithms like maximum ratio combining (MRC) or maximum-likelihood (ML) detection.

An additional antenna-specific coding may be applied to the signals before transmission to further improve the robustness. This is illustrated in the following example.

Example 10.2: In a MIMO system with two transmit antennas and one receive antenna, compare the receiver SNR and data rate of the following two transmit diversity schemes. The receiver uses MRC and the data rate is given by $S = \log_2(1 + SNR)$. The white Gaussian noise has power σ_0. Assume the channel is static over time.

1) In each time slot, each antenna transmits the same symbol s with power $p/2$.
2) Use Alamouti's code. In the first time slot, the first antenna transmits s_1 and the second antenna transmits s_2. In the second time slot, the first antenna sends $-s_2^*$ and the second antenna sends s_1^*. All symbols are transmitted with power $p/2$.

Solution: 1) The received signal in each time slot is

$$y = \sum_{i=1}^{2} h_{1i}s + n = hs + n, \tag{10.9}$$

where $h = h_{11} + h_{12}$. Correspondingly, the SNR is

$$\Gamma_1 = \frac{E\{|h|^2 p/2\}}{\sigma_0} = \frac{(h_{11}^2 + h_{12}^2)p}{2\sigma_0}.$$

The rate is $S_1 = \log_2\left(1 + \frac{(h_{11}^2+h_{12}^2)p}{2\sigma_0}\right)$.

2) With Alamouti's code, the received signal in the first slot is

$$y_1 = h_{11}s_1 + h_{12}s_2 + n_1.$$

The received signal in the second slot is

$$y_2 = -h_{11}s_2^* + h_{12}s_1^* + n_2.$$

The received signals at the two time slots form a vector and the system can be written in the following format:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2^* \end{pmatrix} = \begin{pmatrix} h_{11} & h_{12} \\ h_{12}^* & -h_{11}^* \end{pmatrix} \begin{pmatrix} s_1 \\ s_2 \end{pmatrix} + \begin{pmatrix} n_1 \\ n_2^* \end{pmatrix} \triangleq \mathbf{h} \begin{pmatrix} s_1 \\ s_2 \end{pmatrix} + \mathbf{n}. \tag{10.10}$$

$$\begin{bmatrix} x_1^{(k)} & x_1^{(k+1)} \\ x_2^{(k)} & x_2^{(k+1)} \end{bmatrix} = \overset{\xrightarrow{\text{subcarrier}}}{\begin{bmatrix} S_0 & S_1 \\ -S_1^* & S_0^* \end{bmatrix}} \begin{array}{l} \text{Antenna 1} \\ \text{Antenna 2} \end{array}$$

Figure 10.6 SFBC with two transmit antennas.

With the MRC receiver,

$$\begin{aligned} \mathbf{y}' &= \mathbf{h}^H\mathbf{y} = \mathbf{h}^H\mathbf{h}\begin{pmatrix} s_1 \\ s_2 \end{pmatrix} + \mathbf{h}^H\mathbf{n} \\ &= \begin{pmatrix} |h_{11}|^2+|h_{12}|^2 & 0 \\ 0 & |h_{11}|^2+|h_{12}|^2 \end{pmatrix}\begin{pmatrix} s_1 \\ s_2 \end{pmatrix} + \mathbf{n}', \end{aligned} \tag{10.11}$$

where $\mathbf{n}'$ is zero-mean complex Gaussian with

$$E(\mathbf{n}'\mathbf{n}'^H) = \begin{pmatrix} |h_{11}|^2+|h_{12}|^2 & 0 \\ 0 & |h_{11}|^2+|h_{12}|^2 \end{pmatrix}\sigma_0.$$

So,

$$y_i' = (|h_{11}|^2+|h_{12}|^2)s_i + n_i', i = 1,2.$$

The SNR for the two symbols is the same,

$$\Gamma_2 = \frac{(h_{11}^2+h_{12}^2)p}{2\sigma_0}.$$

The rate is $S_1 = \log_2\left(1+\frac{(h_{11}^2+h_{12}^2)p}{2\sigma_0}\right)$.

Note: while the rate is the same, the first scheme has no diversity and the second achieves the full order-two diversity in fading channels. The first scheme can be improved to send the same symbol in two consecutive time slots over both antennas. This will achieve the full diversity but the rate will be reduced by half.

In the following we introduce the coding for two or four transmit antennas in LTE systems. When the base station has two transmit antennas, a frequency-based version of the Alamouti code, named space frequency block code (SFBC), is used. The code is designed such that the transmitted streams are orthogonal to each other and the optimal SNR can be achieved with a linear receiver. It can be shown that orthogonal codes only exist when there are two transmit antennas [V. Tarokh et al., 1998]. Assume two symbols S_0 and S_1 are to be sent on two transmit antennas. The symbols transmitted from the two antennas on a pair of adjacent subcarriers are shown in Figure 10.6, where $x_i^{(k)}$ denotes the symbol transmitted on the kth subcarrier from the ith antenna. X^* is the complex conjugate of X.

When the base station has four antennas, a combination of SFBC and frequency switched transmit diversity (FSTD) is used since no orthogonal codes exist for more than two antennas. With FSTD, different antennas use different sets of subcarriers. In

$$\begin{pmatrix} x_1^{(k)} & x_1^{(k+1)} & x_1^{(k+2)} & x_1^{(k+3)} \\ x_2^{(k)} & x_2^{(k+1)} & x_2^{(k+2)} & x_2^{(k+3)} \\ x_3^{(k)} & x_3^{(k+1)} & x_3^{(k+2)} & x_3^{(k+3)} \\ x_4^{(k)} & x_4^{(k+1)} & x_4^{(k+2)} & x_4^{(k+3)} \end{pmatrix} \quad \begin{array}{cl} \xrightarrow{\text{subcarrier}} & \\ \begin{pmatrix} S_0 & S_1 & 0 & 0 \\ 0 & 0 & S_2 & S_3 \\ -S_1^* & S_0^* & 0 & 0 \\ 0 & 0 & -S_3^* & S_2^* \end{pmatrix} & \begin{array}{l} \text{Antenna 1} \\ \text{Antenna 2} \\ \text{Antenna 3} \\ \text{Antenna 4} \end{array} \end{array}$$

Figure 10.7 SFBC + FSTD with four transmit antennas.

LTE, the 2 × 2 SFBC is mapped to four subcarriers of the four antennas as shown in Figure 10.7.

Spatial multiplexing

With spatial multiplexing, different streams of data are sent simultaneously on the same RBs by exploiting the spatial-domain channel characteristics. As discussed previously, the rank of **h** determines the number of linearly independent rows or columns in **h** and therefore the maximum number of independent data streams that can be transmitted simultaneously. When all the data streams are allocated to a single UE, it is called single-user MIMO (SU-MIMO). When the streams are assigned to two or more UEs, it is called multi-user MIMO (MU-MIMO). While SU-MIMO increases the throughput of an individual UE, MU-MIMO improves the overall network capacity.

To have good transmission qualities, the channels of different antennas need to be sufficiently decorrelated. The rank of the channel **h** and its singular values are important in determining whether spatial multiplexing should be employed for good performance. A larger inter-antenna distance reduces the correlation between the channels of different antennas. Therefore the antennas should be placed far enough apart that the spacing between antennas is at least half a carrier wavelength.

10.2 Physical layer for uplink

OFDMA is not favorable for the uplink communications because of its weak peak-to-average power ratio (PAPR) properties, which result in poor uplink coverage and high power consumption of UEs. The uplink transmission desires sufficiently low PAPR to avoid excessive cost and power consumption of power amplifiers. Therefore LTE uses single carrier frequency division multiple access (SC-FDMA) in the uplink, which has better PAPR properties than OFDMA.

10.2.1 Basics of SC-FDMA

Figure 10.8 illustrates the principle of SC-FDMA. The transmission process of SC-FDMA is very similar to OFDMA and the only differences are those in the two dashed squares. At the transmitter side, an *M*-point FFT block and *M*-to-*K* subcarrier mapper are placed in front of the IFFT block. The FFT block transforms

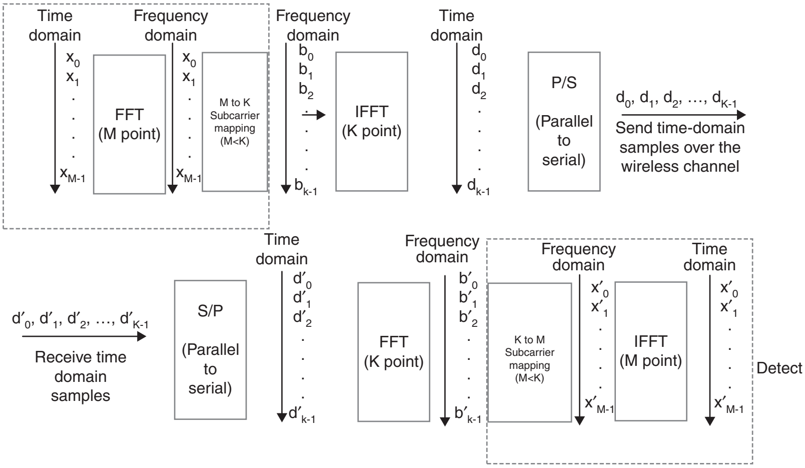

Time domain
x_0 x_1 . . . . x_{M-1}
FFT (M point)
Frequency domain
x_0 x_1 . . . . x_{M-1}
M to K Subcarrier mapping (M<K)
Frequency domain
b_0 b_1 b_2 . . . . b_{k-1}
IFFT (K point)
Time domain
d_0 d_1 d_2 . . . . d_{k-1}
P/S (Parallel to serial)
$d_0, d_1, d_2, \ldots, d_{K-1}$
Send time-domain samples over the wireless channel
$d'_0, d'_1, d'_2, \ldots, d'_{K-1}$
Receive time domain samples
S/P (Parallel to serial)
Time domain
d'_0 d'_1 d'_2 . . . . d'_{k-1}
FFT (K point)
Frequency domain
b'_0 b'_1 b'_2 . . . . b'_{k-1}
K to M Subcarrier mapping (M<K)
Frequency domain
x'_0 x'_1 . . . . x'_{M-1}
IFFT (M point)
Time domain
x'_0 x'_1 . . . . x'_{M-1}
Detect

Figure 10.8 Block diagram of SC-FDMA.

the modulation symbols into *M* symbols in the frequency domain, which are mapped onto the subcarriers allocated by the base station. In SC-FDMA, different UEs are assigned different sets of non-overlapping subcarriers. Therefore, the subcarriers that are assigned to other UEs will be assigned zero at the transmitter. The following operations are the same as those in OFDMA. The FFT processing is therefore the fundamental difference between SC-FDMA and OFDMA. At the receiver side, the corresponding operations are reversed to obtain the time domain symbols. Note that the multipath distortion is handled in the same way as in OFDM, that is, removal of CP, FFT transformation to the frequency domain, and channel correction on a per-subcarrier basis.

Each subcarrier in an OFDMA system carries information of only one specific modulation symbol. On the other hand, with SC-FDMA, each allocated subcarrier contains information of all transmitted modulation symbols because they have been spread by the FFT transformation over all the allocated subcarriers. Therefore the SC-FDMA signal is single carrier as the subcarriers are not independently modulated. As a result, the PAPR of SC-FDMA in the time domain is lower than OFDMA. Note that this may not necessarily be true in the frequency domain.

The subcarrier allocation determines the network performance. Two common techniques can be used, localized and distributed modes, as illustrated in Figure 10.9. In the localized mode, each UE is assigned a set of adjacent subcarriers for data transmission; in the distributed mode, the subcarriers assigned to a UE are distributed over the whole frequency band. Figure 10.10 demonstrates the subcarrier assignment for three UEs using either the localized or distributed mode. The subcarrier scheduling can be either static or channel dependent. The channel-dependent method, as introduced in Chapter 4, schedules subcarriers to UEs based on their channel frequency response. The distributed mode offers a frequency diversity gain because the symbol transmitted is spread over the system frequency band. Using the channel-dependent scheduling method in the distributed mode only improves the network performance incrementally.

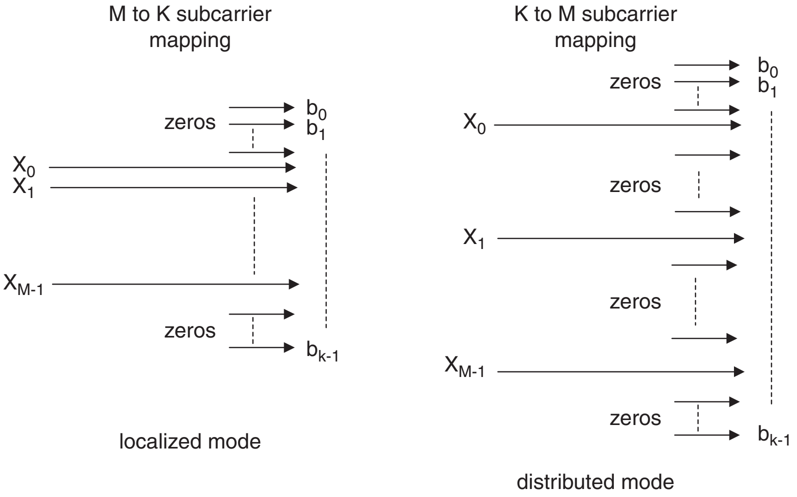


Figure 10.9 Subcarrier mapping modes.

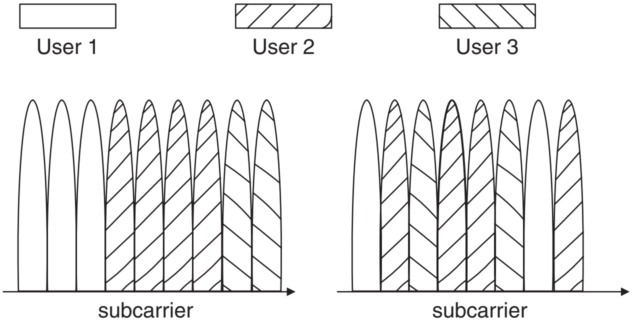


Figure 10.10 Subcarrier assignment for three UEs.

By contrast, channel-dependent scheduling may improve the network throughput significantly for UEs with localized subcarrier mapping as it provides significant multi-user diversity gain.

10.2.2 SC-FDMA parameters for LTE

In this section we briefly explain how SC-FDMA is applied and how radio resources are managed in the LTE uplink transmission.

The LTE uplink uses the same frame structure and resource grid as the downlink, as illustrated in Figures 10.3 and 10.4. Each radio frame has 10 ms, consisting of ten 1 ms subframes. Each subframe has two 0.5 ms slots. As illustrated in the previous section, SC-FDMA uses the same fundamental processing as OFDM, and the same 15 kHz subcarrier spacing is used in both downlink and uplink. The uplink resources are managed in the frequency domain, i.e. before the IFFT. A resource element is also the smallest unit of resource. As for the downlink, there are 12 REs in each PRB, as illustrated in Figure 10.4. The LTE uplink also supports scalable system bandwidths from approximately 1.4 MHz up to 20 MHz, as illustrated in Table 10.1.

The base station schedules uplink resources, assigns PRBs to UEs and informs them about the transmission formats to use. The minimum scheduling interval or transmission time interval (TTI), determined by the subframe length, is 1 ms. On the other hand, the uplink is usually limited by the transmission power of UEs, especially those at cell edges. It is likely that a packet cannot be transmitted with an acceptable packet error rate in a subframe. LTE exploits an efficiency technique, called TTI bundling, for improving the uplink coverage. With TTI bundling, a single transport block can be transmitted repeatedly in several consecutive subframes while using only a single set of signaling. In LTE, up to four TTIs can be bundled together to improve cell edge coverage.

Almost all the uplink transmissions use the localized mode to simplify the data transmission. To exploit the frequency diversity and avoid interference, frequency hopping can be used. There are two hopping modes, inter-subframe hopping and inter- and intra-subframe hopping. With inter-subframe hopping, the frequency or subcarriers allocated for data transmission hops every allocated subframe. With inter- and intra-subframe hopping, a frequency hop may occur both between and within subframes. The distributed mode is used only for the uplink sounding reference signals, which are transmitted by UEs and enable the base station to measure the uplink channel and perform uplink frequency-selective scheduling.

10.2.3 LTE random access

In LTE, UEs can only send data when it is synchronized with the base station. The random access channel (RACH) is mainly used by non-synchronized UEs for their initial network access and should not carry any data. Mobile UEs use RACH to achieve uplink synchronization, after which the base station schedules uplink resources for them to send data. There are two forms of random access procedures, contention-based and contention-free, which are illustrated in Figure 10.11 and 10.12 respectively.

Contention-based random access

As shown in Figure 10.11, there are four steps in contention-based random access.

In the first step, the UE selects and transmits one of the $64 - N$ available RACH preambles, where N preambles are reserved by the base station for contention-free random access. The RACH preambles are basically specific patterns of signals. The preamble values differentiate requests from different UEs. If two UEs apply the same RACH preamble at the same time, there can be a collision.

After receiving the preamble, the base station sends a random access response to the UE. If multiple UEs select the same preamble, each of them will receive the response. The response conveys information like timing alignment instruction for synchronization, resource allocation for the transmission in the third step, backoff indicator to instruct the UEs to back off for a period of time before attempting another random access, and so on. The UE expects to receive the response within a time window and will retransmit the preamble if no response is received within the window.

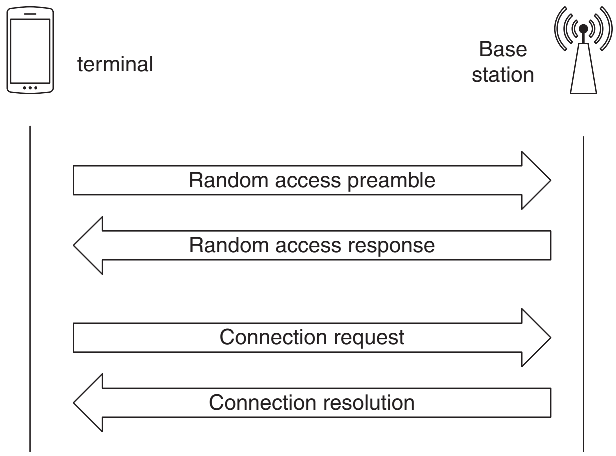


Figure 10.11 Contention-based random access in LTE.

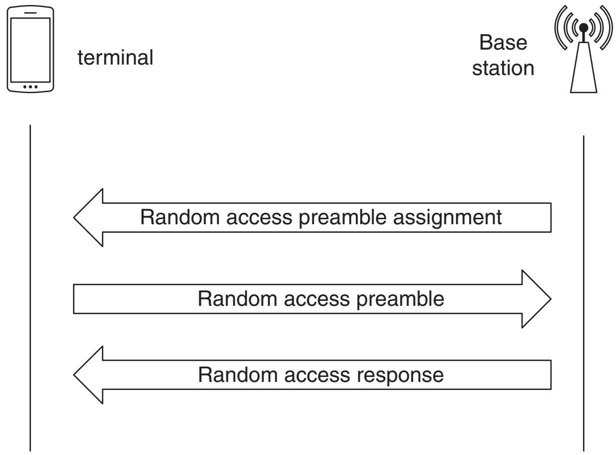


Figure 10.12 Contention-free random access in LTE.

It will also increase the transmission power of the preamble by a fixed step for each retransmitted preamble.

In the third step, the UE sends a connection request to the base station. This is the first scheduled uplink data transmission and hybrid automatic repeat request (HARQ) is used. The request contains the actual random access message, e.g. connection request, scheduling request, and so on. If multiple UEs have selected the same preamble in the first step, a collision occurs. The UEs will also collide in the same uplink time–frequency resources when transmitting their messages in the third step. If no step 3 messages can be decoded for any UE, the UEs will restart their random access procedures after the maximum number of HARQ retransmissions is reached. If the message of one UE is decoded successfully in step 3, the contention for the other UEs is still unresolved and will be resolved in the following step.

In the last step, the base station responds with the contention resolution message. In case of a collision, the HARQ feedback is only sent to the UE with successful decoding in step 3. Other UEs which do not receive the message in step 4 from the base station will understand there was a collision and will quit the current random access and start another one later by random backoff.

The UE that receives the message in the fourth step will move to the following data transmission procedures.

Contention-free random access

In some cases the delay during the random access needs to be as small as possible. This can for example be during a handover or resumption of existing traffic. Contention-free random access can be used in these situations. In this type of random access, a dedicated preamble is allocated to the UE on a per-need basis. The procedure is shown in Figure 10.12. A reserved preamble will only be used by the UE that the preamble was assigned to and steps 3 and 4 in the contention-based random access are not needed.

10.3 Interference management in LTE

The LTE system is envisioned to provide high data rates to support emerging mobile broadband applications. For this, it tries to fully utilize available bandwidth. Therefore, inter-cell interference management is of profound importance to the LTE system. Plenty of interference management concepts for LTE have been proposed, which span from simple interference avoidance to sophisticated interference cancellation. In this section, we will have a look at two distinct interference management techniques.

10.3.1 Soft frequency reuse

Let us start with a rather simple interference avoidance scheme. It has been found difficult to apply the loose frequency reuse of the voice era to the LTE system because it seriously reduces the amount of resource available to each cell. Universal frequency reuse ($K = 1$) is preferred in order to utilize the radio resource as aggressively as possible. However, this in turn leads to a significant performance degradation, especially when the UE approaches the cell edge. Thus, tradeoff between the overall system throughput and the performance of cell edge mobiles has become one of the most important design challenges in the LTE system. The idea of reuse partitioning (recall Section 7.2.1) is regarded as a good compromise between them, and has attracted researchers.

As discussed earlier, LTE employs a physical layer structure based on OFDMA which is very flexible in allocating transmission power to resource blocks in the frequency and time domain. The concept of frequency channel in traditional voice cellular system has been converted to the resource blocks that are allotted different power and modulation levels. Several variants of reuse partitioning have been proposed that exploit the flexibility of OFDMA. Soft frequency reuse (SFR), which was originally proposed in [WG1 3GPP TSG-RAN, 2005], is one of them.

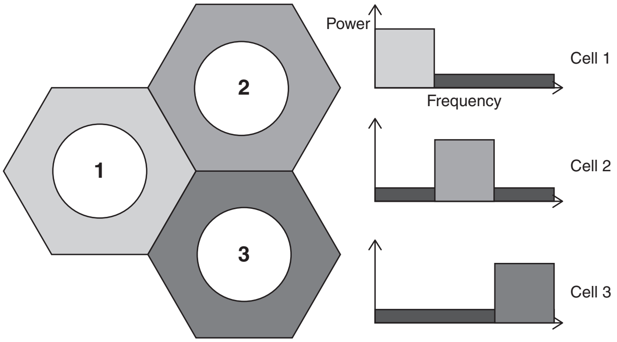


Figure 10.13 Concept of soft frequency reuse.

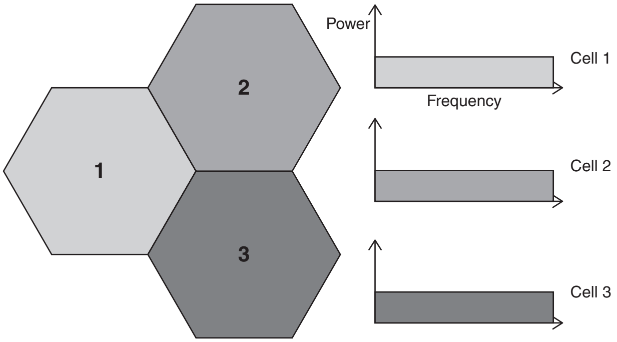


Figure 10.14 Concept of universal frequency reuse ($K = 1$).

In SFR, the cell area is divided into two: cell-center and cell-edge. Then a cluster size of 3 ($K = 3$) is applied to the cell-edge, whereas $K = 1$ is adopted by the cell-center. Different transmission powers are allocated to the cell-center and cell-edge such that the power for the cell-edge is higher than that for the cell-center. This enables the resource allocated to the cell-edge in a cell (namely the cell-edge band) to be reused by the neighboring cell for the cell-center (cell-center band). Therefore, higher resource utilization than the conventional reuse partitioning scheme is available. Resource allocation to the UEs in the cell-edge area are restricted to the cell-edge band in order to ensure a desired SIR level, but the UEs in the cell-center can use both bands. Figure 10.13 illustrates the basic concept of SFR, which you can compare with the illustrations of universal frequency reuse and the cluster size of 3 in Figure 10.14 and Figure 10.15 respectively.

Example 10.3: Let us examine the ratio of transmission power between the cell-center band and cell-edge band. Let P_t be the total transmission power of a base station. Also, let $P_{t,c}$ and $P_{t,e}$ be the transmission power per resource block allocated to the cell-center and cell-edge respectively. The number of resource blocks available at the base station in a timeslot is denoted by N_{RB}. The cell-edge band employs a power amplification

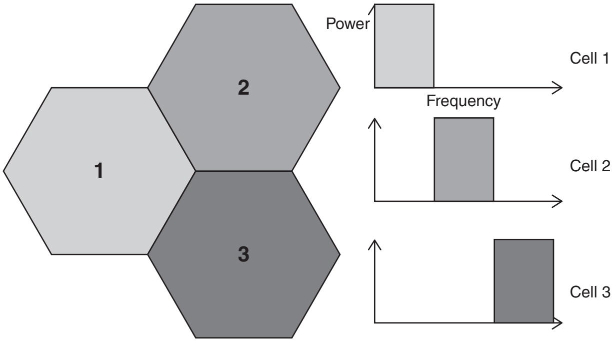


Figure 10.15 Concept of cluster size of 3 ($K = 3$).

factor of a, that is, $P_{t,e} = aP_t/N_{RB}$. Suppose the cell-edge band accounts for 1/3 of the whole resource blocks. What is the ratio between $P_{t,c}$ and $P_{t,e}$?

Solution: Since the total transmission power must remain constant, the following relationship holds:

$$N_{RB}\left(\frac{2}{3}P_{t,c} + \frac{1}{3}P_{t,e}\right) = P_t. \tag{10.12}$$

Simple calculation gives

$$\frac{P_{t,c}}{P_{t,e}} = \frac{3-a}{2a}. \tag{10.13}$$

Let us assume that power amplification of 3 dB is applied to the cell-edge band, i.e., $a = 2$. This results in a power ratio of 0.25, which means a 3 dB decrease in the transmission power of the cell-center band resulting in a 6 dB difference from the power of the cell-edge band.

Example 10.4: In this example, we will see how the performance of a UE varies depending on its location under different frequency reuse configurations: universal frequency reuse ($K = 1$), cluster size of 3 ($K = 3$), and SFR. Let us consider a hexagonal cell surrounded by two tiers of homogeneous interfering cells. A UE in the target cell is moving towards the cell border as illustrated in Figure 10.16. The approximate formula for SIR in Chapter 7 cannot be used in this example because the reuse distance is not large enough with $K = 1$ and $K = 3$. Thus we will calculate the signal and interference power of the UE as a function of its location. For simplicity, background noise and fading effects are ignored.

Let x be the distance between the UE and the serving base station. By assuming a path loss exponent of 4, the received signal power is $P_t x^{-4}$. The distance between the mobile and the neighbor base station j is denoted by $D_j(x)$. For the case of $K = 1$, the interference comes from all the other cells with the same transmission power. Hence,

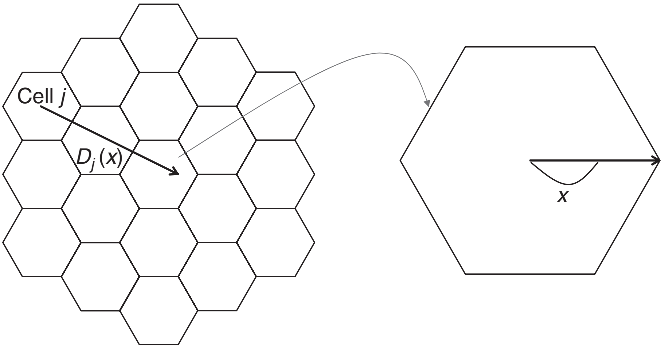


Figure 10.16 An illustration of a UE moving towards the cell border surrounded by two-tier interfering cells.

the SIR is given by

$$\Gamma_{K1}(x) = \frac{x^{-4}}{\sum_{j\in\Psi} D_j(x)^{-4}}, \tag{10.14}$$

where Ψ is the set of surrounding base stations.

Now consider the case of $K = 3$. The cells are grouped into three reuse clusters. Let Ψ_k be the set of base stations in cluster k. Assume that the serving cell belongs to Ψ_1 without loss of generality. Then the SIR of the UE under the reuse factor 3 is as follows:

$$\Gamma_{K3}(x) = \frac{x^{-4}}{\sum_{j\in\Psi_1} D_j(x)^{-4}}. \tag{10.15}$$

When SFR is applied to the system, the cell-center and cell-edge bands experience different SIRs. We still assume the serving cell belongs to Ψ_1. In the cell-center band, the signal power is reduced, whereas the UE receives the strengthened signal in the cell-edge area as discussed in Example 10.3. Interference comes from all surrounding cells in both cases. However, the UE has the strongest interferers in its vicinity for the case of the cell-center band because some interfering cells in the first tier use the same resource block as their cell-edge band with amplified power. In the cell-edge band, there is no strong interferer in the first tier. The SIR of the cell-center and cell-edge bands under SFR is

$$\Gamma_{center}(x) = \frac{P_{t,c}x^{-4}}{\sum_{j\in\Psi_2} P_{t,e}D_j(x)^{-4} + \sum_{j\in\Psi_1,\Psi_3} P_{t,c}D_j(x)^{-4}}, \tag{10.16}$$

$$\Gamma_{edge}(x) = \frac{P_{t,e}x^{-4}}{\sum_{j\in\Psi_1} P_{t,e}D_j(x)^{-4} + \sum_{j\in\Psi_2,\Psi_3} P_{t,c}D_j(x)^{-4}}. \tag{10.17}$$

Figure 10.17 depicts the SIR of the UE as a function of normalized x for the cell radius D_o. It is observed that $\Gamma_{center}(x)$ is lower than $\Gamma_{K1}(x)$ due to the reduced signal strength, and $\Gamma_{edge}(x)$ is lower than $\Gamma_{K3}(x)$ due to the increasing amount of interference. Although the SFR does not bring about an improved SIR compared with the conventional frequency reuse schemes ($K = 1$ and $K = 3$), it is useful for striking a balance between the system throughput and the edge UE performance. Assume the system is fully loaded, and the UE of interest can claim all the available resources in

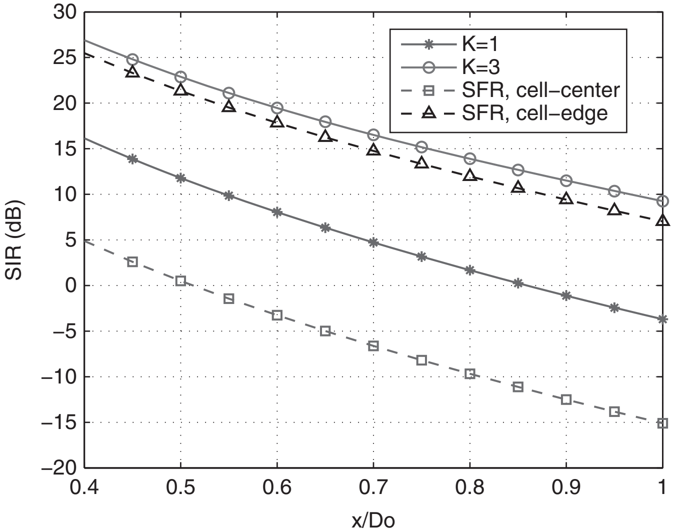


Figure 10.17 SIR of a mobile as a function of the distance from the serving base station.

its serving cell. Using Shannon's formula, the achievable data rate of the UE at x under $K = 1$ is given by

$$r_{K1}(x) = W\log_2(1 + \Gamma_{K1}(x)), \tag{10.18}$$

where W denotes the bandwidth of the system. With $K = 3$, the bandwidth must be reduced by 1/3. Thus,

$$r_{K3}(x) = \frac{1}{3}W\log_2(1 + \Gamma_{K3}(x)). \tag{10.19}$$

For the case of SFR, we assume that the radius of the cell-center area corresponds to $0.7D_o$. Recall that the UE can utilize all the resources (i.e., both cell-center band and cell-edge band) if it is within the cell-center, whereas it can only employ the cell-edge band in the outside of the cell. Then the capacity in the inner region is

$$r_{center}(x) = \frac{2}{3}W\log_2(1 + \Gamma_{center}(x)) + \frac{1}{3}W\log_2(1 + \Gamma_{edge}(x)), \tag{10.20}$$

and the capacity in the outer region is

$$r_{edge}(x) = \frac{1}{3}W\log_2(1 + \Gamma_{edge}(x)). \tag{10.21}$$

As shown in Figure 10.18, $K = 1$ provides the highest capacity when the UE stays close to the cell center. However, the achievable capacity plummets as the UE approaches the cell boundary. $K = 3$ offers relatively even performance across the cell. This is good for the UEs at the cell edge, but the UEs in the good circumstances are also deprived of a high data rate. SFR falls into the middle of the two conventional schemes, striking a good balance between high capacity in the inner cell and quality of service in the outer cell.

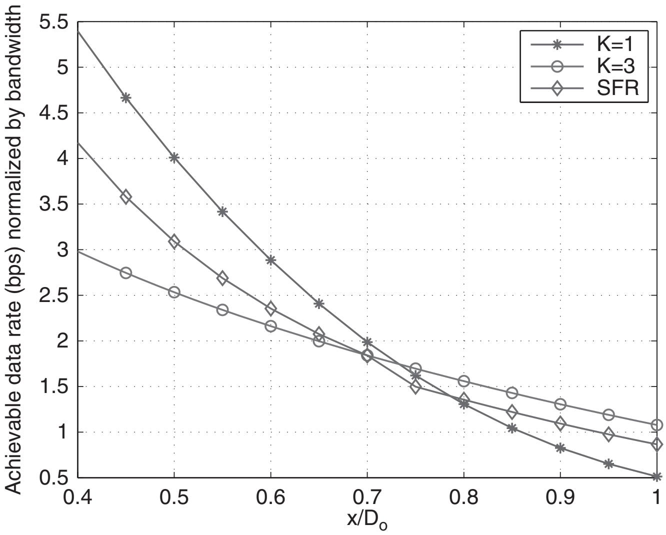


Figure 10.18 Achievable capacity of a mobile under different frequency reuse schemes.

10.3.2 Coordinated multi-point transmission

We learned in Chapter 7 that adaptivity to traffic and wireless channels can improve the system performance significantly. Interference management can be much more efficient if the instantaneous traffic and channel information is shared among the neighboring cells and exploited. The base stations can *cooperate* with each other for better control of the interference. Various levels of multi-cell cooperation are envisaged depending on the extent of available information and processing power. In an ideal case, where the base stations perfectly share all information (precise channel gain matrix, modulation scheme, and the data stream to the UEs) and make an optimal decision, they can act as a single base station entity with a multiple antenna array that is geographically separated, completely removing inter-cell interference and even helping each other.

Multi-cell cooperation requires frequent information exchange between base stations and timely decisions for the group of cells. The capacity and delay of the backhaul connection between base stations has been the major barrier to real-time cooperation. The capability of joint signal processing has also been a major challenge. The recent advances in microprocessors and the growing use of optical fiber backbone have made multi-cell cooperation nearly practicable not only in laboratory environments but also in the field.

In the framework of LTE, coordination between multiple base stations, namely *coordinated multi-point transmission* (CoMP), is under active investigation. CoMP includes various levels of multi-cell cooperation techniques ranging from adaptive interference avoidance (i.e., multi-cell scheduling) to cooperative interference cancellation. A *cell cluster* or *bunch* is used as the unit of CoMP operation because thousands of base stations are usually deployed by an operator to provide seamless coverage, and thus it is almost impossible for massive base stations to cooperate simultaneously. Hence, two or more base stations form a cluster within which the CoMP operation is performed.

The basic concepts of CoMP schemes are described below. Here the CoMP concepts are classified into coordinated scheduling, coordinated beamforming, and joint transmission.

- **Coordinated scheduling**: One entity assumes the scheduling of all the cells in the cluster. The decision can be made by one of the base stations in the cluster or a higher entity. In order to implement coordinated scheduling, the decision-maker (scheduling entity) must have the relevant information about all the mobiles involved in the scheduling process, which usually consists of the channel state information (CSI) of the mobiles and the urgency of the data. With coordinated scheduling, any UE is served only by one base station at a time. Therefore, the data stream does not need to be exchanged between the base stations. The scheduling metric and decision should, however, be exchanged. This requires a reliable and low-latency backhaul connection.
- **Coordinated beamforming**: When the precise CSI of the links between the base stations and UEs is available at the decision-maker, it can not only schedule the UEs to avoid interference but also precode the transmitted signals to cancel the interference from the different cells. Similar to the previous case, a UE is served only by one base station. Therefore, high capacity backhaul is not needed. However, the sharing of the precise CSI matrix requires a very fast and reliable backhaul connection. The burden of channel estimation and reporting at the UE side also increases.
- **Joint transmission**: If the backhaul connection improves further such that the base stations are capable of sharing their data stream to be transmitted, joint transmission of the data by multiple base stations is possible. When it is difficult to achieve perfect synchronization of base stations, joint transmission can be used to combine the signal strengths from the base stations. This resembles soft handover in CDMA systems. With tight time-and-phase synchronization, the cluster of base stations can be regarded as a single base station with a multiple antenna array. Then, the joint transmission is equivalent to a single-cell MU-MIMO technique with geographically large antenna separation.

It is reported that the CoMP techniques have great potential to improve the performance of mobiles, particularly at the cell boundary. However, there are still challenges to be addressed. First, a fast and reliable backhaul connection is a prerequisite for CoMP. Large backhaul capacity is required as well for the implementation of joint transmission. If an investment in the backhaul installation is needed to support the cooperation, the tradeoff between performance and cost should be examined. Second, the CoMP schemes come with an estimation and feedback overhead. More reference symbols should be sent from the base stations so that the UEs in the cluster can estimate the channel gains with the multiple base stations. This reduces the radio resource available for data transmission. Estimation and feedback of channel information for the multiple cells also consumes the energy and radio resources that the UEs have. Third, the CoMP operation is performed within a cluster. Even if the interference is completely cancelled inside the cluster, inter-cluster interference still exists. Including more base stations in the cluster would suppress more interference, but would increase the estimation and feedback overhead at the same time. Thus, it is not trivial to find the optimal cluster size. Dynamic selection of the cluster depending on the radio propagation environment and traffic demand is also an interesting research topic.

Exercises

10.1 In OFDM, the waveform can be generated by using an inverse fast Fourier transform (IFFT), e.g.

$$d(n) = \frac{1}{N}\sum_{k=0}^{N-1} b_k e^{j2\pi kn/N}. \tag{10.22}$$

k is the subcarrier index and n the time index. b_k is the complex symbols modulated onto subcarrier k. Assume the symbol period to be NT_s where T_s is the sample period. What is the subcarrier spacing? What is the frequency of the kth subcarrier?

10.2 Consider a MIMO system with six transmit and four receive antennas. For each of the following channels, what is the multiplexing gain, that is, how many independent data streams can be transmitted reliably?

$$\mathbf{h} = \begin{pmatrix} 1 & 1 & -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & 1 & -1 & 1 \\ 1 & 1 & 1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & 1 & 1 \end{pmatrix}$$

$$\mathbf{h} = \begin{pmatrix} -1 & -1 & -1 & 1 & -1 & -1 \\ 1 & 1 & 1 & 1 & -1 & 1 \\ -1 & 1 & -1 & 1 & -1 & 1 \\ 1 & 1 & 1 & -1 & 1 & 1 \end{pmatrix}$$

10.3 Consider the system in Example 10.2. Determine the receiver SNR and data rate of the following space–time coding scheme. On the first time slot, the first antenna transmits symbol s. On the second time slot, the second antenna transmits the same symbol.

10.4 Partial frequency reuse (PFR) is another variant of reuse partitioning designed for the LTE system. The radio resource is divided into cell-center band and cell-edge band in PFR. The cell-center band is dedicated to the UEs in the inner cell. Transmission power can be amplified in the cell-edge band. Contrary to SFR, the cell-edge band is further partitioned so that each cell utilizes only a portion of the radio spectrum. Figure 10.19 depicts the concept of PFR.

Let a be the amplification factor applied to the cell-edge band, and let b be the portion of spectrum allocated to the cell-edge band.

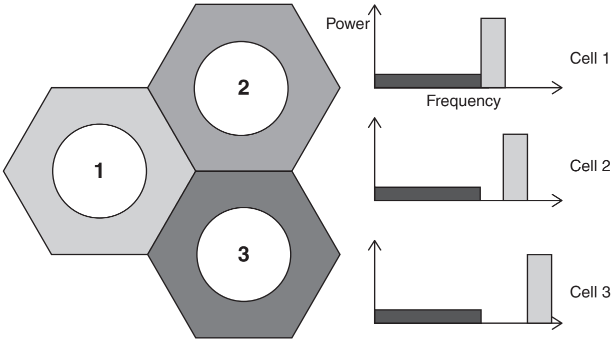


Figure 10.19 Concept of partial frequency reuse.

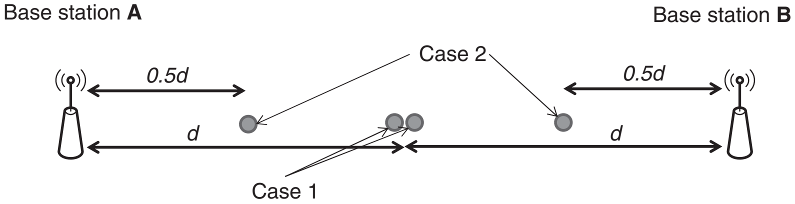


Figure 10.20 Reuse partitioning scheme with two zones.

a) Express the ratio of transmission power per resource block between the cell-center band and cell-edge band.

b) An LTE base station has a transmission power of 43 dBm with a bandwidth of 20 MHz. Spectrum division between the cell-center and cell-edge bands is 2:3. Calculate the power allocated to each band for the amplification factor of 6 dB.

10.5 Consider a one-dimensional LTE system consisting of two cells. The two cells have the same transmission power and antenna gain. Each base station is currently serving one UE. Figure 10.20 illustrates the system model. We will examine three interference coordination schemes as follows:

i) Universal frequency reuse ($K = 1$).

ii) Equal resource division between the two cells ($K = 2$).

iii) Joint transmission with signal power combining: the two cells send the same data to the UE, and it combines the received signal power. The two base stations jointly schedule their UEs such that the same time portion is assigned to each mobile.

Obtain the achievable capacity of the two UEs under the above schemes when

(a) (Case 1) the two mobiles are located at the middle of the base stations.

(b) (Case 2) the mobiles are closer to their serving base station. the ratio of the distance between serving and neighbor base stations is 1:3.

Neglect fading effects. The path loss is assumed to be proportional to the fourth power of the distance. Noise power is assumed to be 3 dB lower than the received signal power from one base station when a mobile is located in the middle.

References

R. W. Chang. 1966. Synthesis of band-limited orthogonal signals for multichannel data transmission. *Bell System Technical Journal*., 45(10), 1775–1796.

R. W. Chang. 1970. *Orthogonal Frequency Multiplex Data Transmission System*. US Patent 3,488,445.

L. J. Cimini. 1985 (July). Analysis and simulation of a digital mobile channel using orthogonal frequency division multiplexing. *IEEE Trans. Commun*., 33(7), 665–675.

S. Stefania, T. Issam and B. Matthew. 2009. *LTE – The UMTS Long Term Evolution: From Theory to Practice*. Chichester: John Wiley and Sons, Ltd.

V. Tarokh, N. Seshadri and A. R. Calderbank. 1998 (Mar.). Space–time codes for high data rate wireless communication: Performance criterion and code construction. *IEEE Trans. Inf. Theory.*, 44(2), 744–765.

WG1 3GPP TSG-RAN. 2005 (May). *Soft Frequency Reuse Scheme for UTRAN LTE*. R1-050507.

11 Wireless infrastructure economics

11.1 Communication infrastructures

Most of this book has covered how to provide effective wireless access services and the efficient utilization of the spectrum resources. In the previous chapters we have also looked at energy aspects. Now we will widen the scope even more as we look at the total resource consumption in a wireless access infrastructure. This will include networks, switches/routers and access ports and the terminals, as described in Chapters 2 and 9. In this chapter, the tradeoff between the resource consumption incurred by adding more access points, i.e. a more expensive infrastructure, and increased capacity and a higher QoS provided to the users, will be reviewed in more detail. We will make comparisons among such diverse quantities as frequency spectrum allocation, equipment, physical infrastructure including towers and antennas, real estate, power consumption, user equipment and maintenance. Such comparisons are naturally made in monetary or economic terms. The total cost of running a wireless access system comprises the cost of all these individual components. In economic terms, this is usually referred to as the *supply side* of operating a network. This cost of providing (supplying) the network has to be compared with revenues that the system provider, for example an operator, can derive from the users, i.e. what the users are willing to pay for the service. The "will" in turn is based on user demand and user satisfaction. Supply and demand are tightly coupled and together determine if operating the network will be a profitable endeavor, i.e. that the revenue will exceed the investment.

Although not always easy, it is usually possible to compute the cost of network operation, as we will see later in the chapter. However, understanding the demand and willingness to pay is difficult to model and will depend on how the services are provided and which actors are providing them. The telecommunications industry has been a global one for over a century, with many actors. Whereas previous chapters have outlined the supply chain of actual communication (transport) services, the picture has been more complicated in the past as the actual content transmitted and the services that have been provided by and through the network have been interwoven with the network service itself. Figure 11.1 gives a rough description of the actors on the scene and their interrelations. In the late 19th century, i.e. in the early days of telephony, companies emerged that provided telephony as a complete end-to-end service to their customers. They provided the wiring, the handsets and the manual switching service. Every telephone company had to provide its own equipment, since

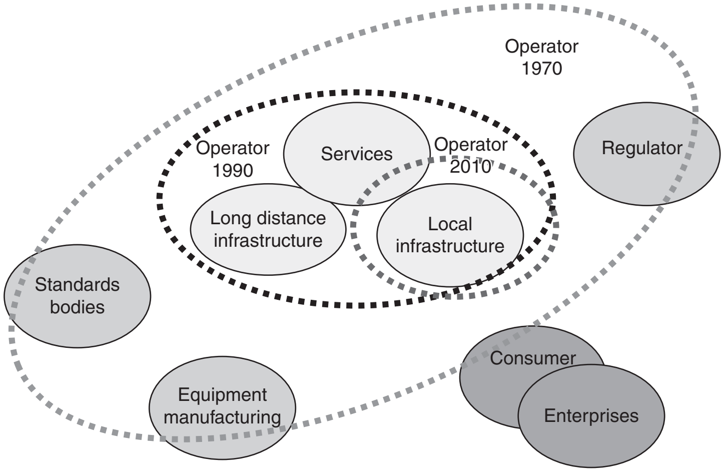


Figure 11.1 Actors and operator roles on the telecommunication scene over four decades.

equipment was not standardized. Local competition between competing operators in the same neighborhood was fierce as regulation was yet to be imposed. When several competing telcos wired the same neighborhood, it led in many cases to a cost crisis, in particular where the wiring costs were high and the customers few. No single telco achieved sufficient scale advantages to become profitable, which in turn resulted in one of two outcomes: either the companies merged or the local lawmakers granted one of them exclusive rights, an *exclusive license*, to operate a telephony system in some area. In both cases a local monopoly was formed where a single telco was operating all telephony services in the area. In many European countries, this local monopoly evolved into a nationwide one, where all of the telecommunication services in the whole country were offered by one company. One may debate whether telephony services constitute what economists call a *natural monopoly* [W. Sharkey, 1982] or not. This term describes those situations where a single company is capable of providing all the services demanded by the consumers at a lower cost than two or more companies due to economies of scale. In the real world things are more complex. The lawmaker, the *regulator*, which is aiming at protecting public interests, may have other objectives besides ensuring that the service is provided at a minimal cost. Another objective is the *public welfare objective*, i.e. providing a telecommunication service in a fair and affordable way, such that anyone, wherever they may live, has access to the service at a reasonable price. Sometimes this is in regulatory terms referred to as a *Universal Service Obligation* (USO). A consequence of such regulation is that in some places the service is made available at a low price even though the marginal cost of providing it is high. These situations are usually termed *market failures*. Under strict market conditions, no operator would provide the service at those prices in these situations.

The most important tool available to the lawmaker has been the licensing of monopoly rights. The sale of monopoly rights has a long history as a device for raising money for the government or monarch of the state. This may have little to do with protecting the public interest, but it is anyway an important element in financing the development of countries. A licensing regime may be used to create a (natural) monopoly where costs are lower. In telecommunications we may achieve benefits such as

- reducing the interference from uncontrolled use of the electromagnetic spectrum,
- public service obligations (USO's), creating services which would not emerge from market processes, and
- imposing quality and standards.

Granting licenses can be done in many ways. One way is to use a market and auction off licenses to the highest bidder. Another is to use beauty contests where licensees are chosen based on the trust the regulating body puts in the licensees' ability to fulfill the license agreement.

The national monopolies were intact in most countries for most of the 20th century. Regulation and telecommunication policymaking is a delicate matter. The intention of the regulator was to apply controls and incentives to make the monopolistic operator behave as if operating in a fully competitive market. This is, however, not straightforward. The operators would adjust themselves to the rules and the result might not always be what the regulator originally intended. The monopolist is required to provide all the services, both the profitable ones and the unprofitable ones, at equitable prices, which on top of everything else might be tightly regulated. With the evolution of technology, many severe discrepancies between the cost of providing the service and the prices charged were laid in the open. The most obvious example may have been the transformation of long distance telephony in the early 2000s. In the early days, providing the wiring for long distance telephony was expensive and the cost was virtually proportional to the distance. With today's technology, a single physical (optical) cable and all of the required routing and switching equipment are shared by thousands of simultaneous voice and Internet connections such that the cost is almost independent of the distance. The local access network, on the other hand, did not change very much in this time of transition. The individual consumer was still to be reached by a twisted pair of wires, or fiber, laid down for his/her exclusive use. With the evolution of technology, advanced network elements, for example switches, routers and so on, have become less and less expensive, whereas the cost of wiring, in particular for local access, is dominated by labor cost, i.e. it increases over time. This can be illustrated with the following simple numerical example.

Example 11.1: Consider two cities, A and B, where we would like to provide network connectivity between their households (consumers), i.e. both local and long distance connecting the users in both cities. Both cities have 100,000 households and the traffic to the outside world can be neglected. Assume that the network cost is dominated by the cost of providing the wiring and that this cost is proportional to the length of the cables. Estimate the relation between the cost of the access network, i.e. last 100 m to the consumer, and the inter-city network.

Solution: The total cost of the wiring in the access network can be computed as

$$C_A = 2 \cdot 100000 \cdot c \cdot 100 = 2 \cdot 10^7 c. \tag{11.1}$$

In a similar way, the inter-city network wiring cost becomes

$$C_I = c \cdot 100 \cdot 10^3 = 10^5 c. \tag{11.2}$$

The ratio between these cost items becomes:

$$\frac{C_A}{C_I} = \frac{2 \cdot 10^7 c}{10^5 c} = 200. \tag{11.3}$$

The means that the cost of the access network is 200 times higher than the cost for the inter-city connections. Although the inter-city connection in itself is 1000 times more expensive than each individual access connection, its cost is shared by the 200,000 consumers.

Even though Example 11.1 is somewhat exaggerated, it still illustrates the key points. Nowadays long distance communication is cheap to provide and local communication is expensive, whereas pre-2000 prices indicated the opposite. This made long distance telephony the first target of emerging new telcos as soon as the licensing regulations were relaxed in the 1980s. As the wave of deregulation and competition swept over the world, most of the other monopoly telcos' business areas have become business opportunities for the incumbent operators. The local infrastructure, which the consumer has learned to view as a public utility, provided virtually free of charge, has for natural reasons had a very low priority among the new actors and very few new investments have been made in this area. Existing or former monopoly operators for many years controlled an extensive access infrastructure with huge sunk investment costs, effectively preventing market entry for new actors at any significant scale for more than 50 years in most western countries. It is only in the recent decade that we see significant deployment of new infrastructure for local access – first cable TV and finally now optical fiber networks. Although efficient reuse of the existing copper-wire infrastructure resisted this development for many years through more and more sophisticated xDSL technology, fiber access solutions now emerge as the only viable solution for broadband ($\leq$ 100 Mbps) access.

Returning to Figure 11.1, we see that the big deregulation in the 1980s shrank the scope of the operator business, as illustrated by the outer oval representing the situation in 1970 and before and the inner oval representing the situation in the early 1990s. But the evolution did not stop there. The next blow to the operator business was the massive deployment of optical fibers for long distance Internet communications, which reduced the cost of long distance communications to almost zero. The availability of IP networks accessible to the end customers had two major consequences: It opened up Internet telephony, e.g. Skype, which created effective competition in the long distance telephony market, the former cash cow of the traditional telecom operator. The other, even more serious consequence, is that all the services that were previous intimately connected to the telecom operators could now be provided by anyone using the IP connectivity. The service monopoly seemingly disappeared overnight and an abundance of services are now provided over the top without any involvement of the operator. The traditional operator is now squeezed to basically being a local access provider (Figure 11.1). From a business perspective this is a serious matter for an operator as the global service providers, for example Google, Facebook, Netflix, generate huge amounts of traffic in the network, although they do not pay for the communication services. Every contribution to cover the cost of the operators has currently been derived from end-user charges. In the next section we will look more closely into the economics of wireless access systems.

11.2 Wireless access economics

The emerging cellular technologies at the beginning of the 1980s represented a milestone that marked significant changes in the telecommunication infrastructure business. In the decades before, we witnessed several breakthroughs in telecommunication technology: optical fiber communications, satellite communications and, not the least, the digital computer. These new technologies mainly improved the economics of long distance communications. Meanwhile, the access networks, still dominated by copper wire technology, were up to that point hardly affected by these changes. Wireless technology, and in particular cellular telephony, however, represented a game changer. This technology had the potential of not only offering market entry opportunities for new players with significantly lower infrastructure investments, but it also provided an additional functionality, mobility. In this section we will investigate the economic driving forces behind the unprecedented success of mobile cellular telephony in somewhat more detail.

In order to do this, start by reviewing Equation (5.13). For the sake of simplicity we will assume that a fixed transmission format is used and that the required QoS is given. The system capacity can in this case be written as

$$\omega \approx c'(\alpha)\frac{C\varpi_\eta}{\gamma_0^{2/\alpha}A_c} = c''\frac{W_{sys}}{A_c} = c''\frac{W_{sys}}{A_{tot}}N_{AP} \text{ (erlangs/km}^2\text{)}, \tag{11.4}$$

where A_{tot} is the area to be covered by the total system and N_{AP} is the number of uniformly distributed access points. The number of channels C is directly proportional to the available system bandwidth, W_{sys}. For the mobile data case the relationship is basically the same,

$$R_A \triangleq \frac{R_{tot}}{A_{tot}} = \frac{\eta\overline{R}(W_{sys})}{A_{tot}}N_{AP} \text{ (Mbps/km}^2\text{)}, \tag{11.5}$$

where $\overline{R}$ is the Shannon cell capacity for a single cell in Equation (5.21), and $\eta\overline{R}$ is the actual data rate achieved in the cell due to interference and implementation losses. $\eta < 1$ can be seen as the relative efficiency of the modulation and RRM schemes. From the same equation we see that $\overline{R}(W_{sys})$ is also directly proportional to W_{sys} in a single channel system, when SINR constraints rarely allow us to use the maximal data rate R_{max}. Of course, when multiple channels are used in each cell in a mobile data system, the rate is also directly proportional to the number of channels, i.e. to the system bandwidth as in Equation (11.4). For a given fixed channel bandwidth W_{ch} we have

$$R_A = \frac{\eta \overline{R}(W_{ch})}{W_{ch} A_{tot}} W_{sys} N_{AP} = \frac{\eta \overline{S}}{A_{tot}} W_{sys} N_{AP} \text{ (Mbps/km}^2\text{)}, \tag{11.6}$$

where we have introduced the cell spectral efficiency

$$\overline{S} \triangleq \frac{\overline{R}(W_{ch})}{W_{ch}}. \tag{11.7}$$

What is most important in the equations above is that the number of users that can be served per area unit (or the total data rate provided for all users per area unit) is directly proportional to the number of access points. The cost of providing the infrastructure includes the cost of providing the base station equipment, base station sites, antennas, towers and so forth, as well as the fixed network connecting the base stations. In the following analysis, we will focus on the operator perspective and we therefore will not consider the cost of the terminals. In the simplest cost model it is assumed that the cost of building and maintaining the infrastructure is simply a linear function of the number of base stations, i.e.,

$$\text{Cost model I: } C_{infra} = c_1 + c_2 N_{AP}. \tag{11.8}$$

The key argument why this assumption should be valid is given by Example 11.1. In fact the approximation becomes better and better as the base station density increases. Each base station is shared by a few users, whereas the core network equipment (routers, switches) are shared by thousands of users. Now, combining Equation (11.4) and Equation (11.6) respectively with Equation (11.8), we get

$$C_{infra} = c_1 + c_2{}' \frac{\omega A_{tot}}{W_{sys}}, \tag{11.9}$$

$$C_{infra} = c_1 + c_2{}'' \frac{R_A A_{tot}}{\eta \overline{S} W_{sys}} = c_1 + c_2{}'' \frac{R_{tot}}{\eta \overline{S} W_{sys}}. \tag{11.10}$$

Now finding the relative cost per user or per transmitted bit yields:

$$\frac{C_{infra}}{\omega A_{tot}} \prec \frac{c_1 + c_2 N_{AP}}{\omega A_{tot}} = c_2{}' + \frac{c_1{}'}{\omega A_{tot}}, \tag{11.11}$$

$$\frac{C_{infra}}{R_{tot}} \prec \frac{c_1 + c_2 N_{AP}}{R_{tot}} = c_2{}' + \frac{c_1{}'}{N_{user} R_{user}}. \tag{11.12}$$

Whenever in a wireless system the fixed cost of providing rudimentary coverage c_1 is small and/or the capacity demand is high, the second term in the two equations above can be neglected and the cost per served subscriber or per bit transferred is almost constant. This is in contrast to fixed access infrastructure, which has a high initial cost to provide any service at all, whereas the marginal cost of transmitting yet another Mbps is usually small.

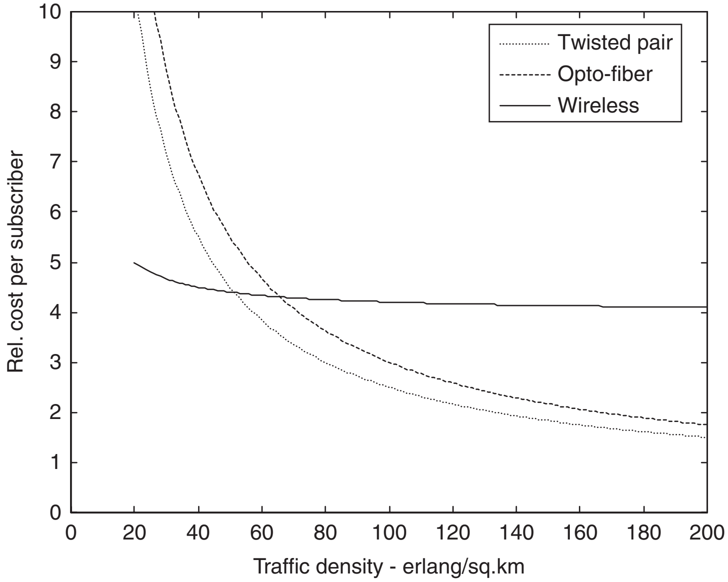


Figure 11.2 Cost for access per subscriber using various access techniques.

An alternative explanation is that in a wireless system there is virtually no equipment or parts that are for the exclusive use of a single user. All infrastructure is shared, which is illustrated in Figure 11.2. We can see that for low and moderate user traffic densities (for example in rural and suburban environments), cellular technology provides the service at the lowest cost. In dense urban environments, such as city centers, where the infrastructure can potentially be shared by many users, the fixed access costs are more favorable. All this aside, consumers would of course also be interested in paying a premium for mobility.

This cost structure that we see in cellular telephony systems turned out to be ideally suited for introducing the technology in the early 1980s. New players on the telecom scene were able to enter the market despite their limited capital base since the new technology required only moderate infrastructure investments. They could start with only a minimal set of base stations barely covering the most densely populated areas and some major highways, and have revenues rolling in. As the demand for the new service was large, the capacity of the many cellular systems was soon exhausted. Since revenues had been substantial, the operators now had the necessary cash to invest in more base stations, increasing the revenues even more, which in turn resulted in increasing investments, all in a fast-moving upward spiral. Payback times for the initial investments of two years were not uncommon, in a business that used to be looking at payback of 30 years or more (copper landline). Although the initial purpose was to provide mobile services, already at the end of the 1990s wireless mobile subscriptions actually surpassed the number of fixed subscribers in several of the pioneering countries. Although tariffs in those days were still substantially higher for wireless mobile services than corresponding fixed services, they became extremely popular in the late 1990s.

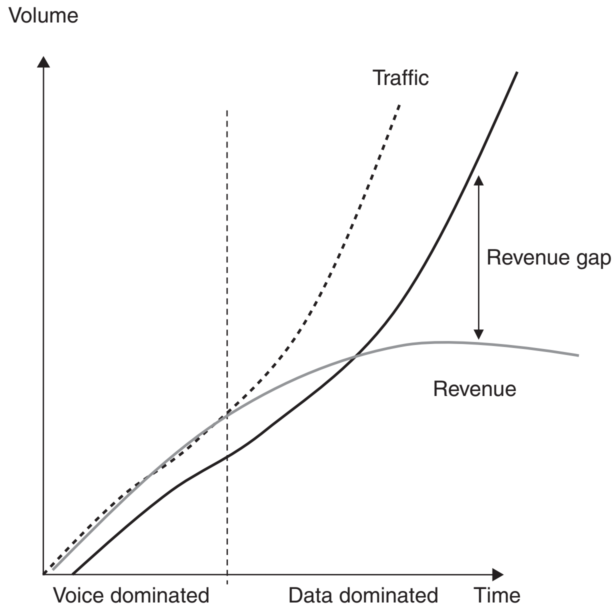


Figure 11.3 Development of data and voice traffic and revenue in wireless access.

At the beginning of the 2000s mobile data entered the scene. The deployment of 3G/UMTS systems suddenly provided many operators with large excess voice capacities. To stimulate the use of mobile data, most operators in 2006–7 started offering a fixed monthly fee for unlimited data packages. These so-called flat rate schemes promoted an Internet-type user behavior, where the marginal click on your web browser is for free. This, in combination with the advent of smartphone apps using cloud-based servers, has caused a very rapid growth of traffic as illustrated in Figure 11.3. In 2007–2012 traffic doubled almost every year and this trend does not show any signs of weakening. The problem is that Equation (11.10) tells us that the cost of the infrastructure providing the capacity to handle this traffic, i.e. the number of required access points, grows linearly with the traffic volume. As in many western countries, now almost everyone already has a phone subscription, there are hardly any new customers that will generate revenue. The revenue that had been steadily increasing due to the per-minute charging regimes of voice traffic, now levels out [J. Zander and P. Mähönen, 2013].

This is caused by the flat rate charging regime, but also by the fact that the increased data capacity makes the voice traffic almost negligible. The price for producing a voice minute is rapidly dropping. As a consequence the revenue derived per bit drops whereas the cost of transmitting that bit is almost a constant. The widening difference, sometimes referred to as the revenue gap, causes significant problems to many operators. Currently, as of 2012, many operators are still able to derive significant revenue from the increasingly profitable per-minute charged voice traffic but alternative technologies, for example voice over IP, and competition has put a significant pressure on voice prices. Inevitably, in the long run, we will need a way to produce mobile data access services at a cost that matches the revenue that can be derived. There are several ways in which this can be achieved:

A. Increase prices for mobile access.
B. Limit the amount of traffic per user.
C. Lower the cost of wireless access.

Alternative A is feasible to some extent only. The growth in the economy is only a few percent/year which means that the users will, in total, not have significantly more money to spend. The addiction to always being connected is, however, strong, and the users will certainly be prepared to spend a somewhat larger share of their budget on wireless access. Alternative B has negative consequences on the user experience, as it will change the habits of the users. C, of course, is the most attractive solution, and in the next section we will dig somewhat deeper into this.

11.3 Spectrum cost and regulation

Although the regulators are easing their grip on the telecommunications market, there are still significant barriers for new actors in the wireless field. One is the strict regulation of the frequency spectrum, another is standardization. In fact these two phenomena are tightly coupled. Although the national states are bound by treaties within the International Telecommunication Union (ITU) prescribing the use of the spectrum, they maintain sovereignty over which national actors should be awarded rights to use the spectrum. Licensing is the method by which national frequency authorities give the licensees exclusive rights to use a certain part of the spectrum for a certain purpose. The regulator seeks, from a public interest perspective, to organize the use of the spectrum in as efficient a manner as possible and to avoid unwanted interference between different services and users. With the increasing traffic in mobile systems, the demand for spectrum has increased. How and to whom to award spectrum licenses is an interesting problem. Methods range from the national regulator, after considering various applications, deciding which actor, for example operator, can make best use of the spectrum, usually known as a beauty contest, to pure market mechanisms such as spectrum auctions. At an auction the bid may be for an unlimited license seeing the spectrum band as a piece of property that later can be resold, or for a time-limited license. The latter type of auction is the most popular among regulators these days, since it gives the opportunity to re-appropriate, or refarm, spectrum when the license period ends. In most licensing agreements the licensee is also required to use equipment and systems that are in compliance with certain standards or technical specifications. In the early days of the ITU (1950s–1970s) the organization also dominated the standardization process. Since the late 1980s the lead roles are now played by the equipment and system manufacturers. Their organizations, such as ETSI (Europe), ARIB (Japan) and TIA (USA), have become the leading standardization bodies.

The value of a piece of frequency spectrum, i.e. how much an operator can be willing to bid in a spectrum auction, can be illustrated by extending the cost model in Equation

(11.8) to

$$\text{Cost model II: } C_{sys} = c_1 + c_2 N_{AP} + c_3 W_{sys}, \tag{11.13}$$

where the constant c_3 corresponds to the licensing fee or cost per MHz. Combining this with Equation (11.10) for large capacities (c_1 negligible) yields:

$$C_{sys} = c_1 + c_2 \frac{R_{tot}}{\eta \bar{S} W_{sys}} + c_3 W_{sys} \approx c_2 \frac{R_{tot}}{\eta \bar{S} W_{sys}} + c_3 W_{sys}. \tag{11.14}$$

The first term in the equation corresponding to the infrastructure cost is decreasing with the amount of spectrum available, whereas the second term is the spectrum cost. Theoretically, if there were no constraints on the available spectrum, the total cost would be minimized by acquiring

$$W^*_{sys} = \sqrt{\frac{c_2 R_{tot}}{\eta \bar{S} c_3}} \text{ (MHz)}. \tag{11.15}$$

At this minimum point the infrastructure cost and the spectrum cost are equal. For a given amount of spectrum W_{sys}, the corresponding spectrum cost when the spectrum is just as expensive as the infrastructure is

$$c^*_3 = c_2 \frac{R_{tot}}{\eta \bar{s} (W_{sys})^2} \text{ (monetary units/MHz)}. \tag{11.16}$$

c^*_3 has the interpretation as the engineering value of a MHz of spectrum. Assume that an operator is interested in providing an additional capacity R_{tot}. He can choose between installing more access points by increasing C_{infra} or buying more spectrum. If the spectrum price is above c^*_3, installing more access points is the best strategy, whereas when the spectrum is less expensive, the operator should invest in more spectrum.

The demand for spectrum has been high and in the rare auctions for licensed spectrum prices have been quite high, which has effectively barred all smaller players from getting access to spectrum. Considerable interest is therefore put into the development of systems with alternative forms of licensing. The most popular form to date is operation in so-called unlicensed frequency bands. In these bands the user is allowed to operate all kinds of equipment and systems fulfilling minimal technical requirements. The regulator provides no guarantees regarding interference from other legal users of the band. Examples of such systems are WLAN systems, for example, as specified by the IEEE 802.11.

11.4 Affordable wideband wireless access

As discussed in the previous section, traditional mobile communication systems that are primarily designed to provide cost efficient wide area coverage for users with moderate bandwidth demands have had a very successful evolution. The charging regime that provided revenues proportional to the traffic demand created an upward spiral, providing systems with more and more capacity at constant or even decreasing cost for the users. It has, however, become clear that flat rate data access causes the revenue gap problem, as revenues basically stay constant whereas the capacity requirements keep increasing. To sustain this type of business the cost per transmitted bit has to go down significantly, in the same way as optical fiber has dropped the communication costs in wireline systems. In this section we will investigate how this may be achieved. We will start by refining the models provided in Section 11.2 and present a simple analysis of the cost structure of two types of access architectures—a universal coverage scenario offering wideband services at all locations and a hotspot scenario where full-rate service is offered only in limited geographical areas. Both infrastructural costs and spectrum licensing costs are included in the analysis.

Consider as previously a wireless access network for mobile data covering a certain service area, A_{tot}. In this area N_{AP} wireless access ports are dispersed. A total of W_{sys} bandwidth is available. Further assume that the users are provided the average bandwidth $\overline{R}$ according to Equation (5.21). We make the simplifying assumption that

$$\overline{R} \approx b \cdot R_{min}, \tag{11.17}$$

where R_{min} is the minimum data rate and $b > 1$ is a constant.

As in Chapter 5, the mobiles use a transmitter power of at most P_{batt}, giving them a maximal range of D_0 where at least R_{min} can be provided (Figure 11.4). This corresponds to the maximal distance where the signal-to-noise requirement can be met in the absence of interference from other users in the system, where

$$\begin{aligned} R_{min} &= cW_{ch} \log_2 (1 + \Gamma(D_0)) \\ &= c\frac{W_{sys}}{K} \log_2 \left(1 + \frac{aP_{batt}}{N{D_0}^{\alpha}}\right). \end{aligned} \tag{11.18}$$

We assume that the path loss is proportional to the αth power of the distance, where α is typically in the range 2–4 ($\alpha = 2$ corresponds to line-of-sight communication paths). Since the received power required to maintain a certain signal-to-noise ratio

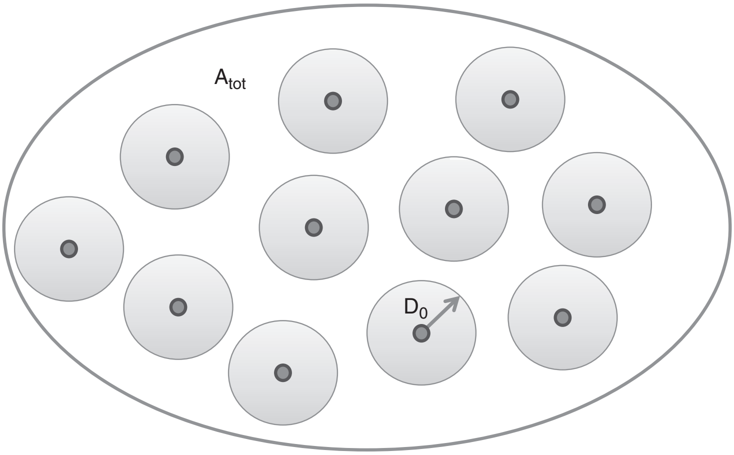


Figure 11.4 Service area: partial coverage and notation.

is proportional to the user bandwidth:

$$\overline{R} = c' \frac{W_{sys}}{K} \log_2\left(1 + \frac{aP_{batt}}{ND_0{}^{\alpha}}\right), \tag{11.19}$$

$$D_0 = \sqrt{\frac{\alpha P_{batt}}{N\left(2^{\frac{\overline{R}K}{c'W_{sys}}} - 1\right)}}, \tag{11.20}$$

where c' is a constant dependent on the actual transmission system, for example modulation, coding and so on. Consider a system where each base station provides access service to the mobile terminals in its immediate surroundings. The service areas of the base stations are approximately circles (cells) of the same radius. We can now calculate the cell area,

$$A_{cell} \approx \pi D_0{}^2. \tag{11.21}$$

For equidistantly dispersed base stations, we could achieve seamless or full coverage when

$$N_{AP}A_{cell} \approx N_{AP}\pi D_0{}^2 > A_{tot}, \tag{11.22}$$

$$N_{AP} > \frac{A_{tot}}{\pi D_0{}^2} = N_{AP}{}^*, \tag{11.23}$$

i.e. when the maximum base station range D_0 exceeds the critical cell radius D^* given as

$$D_0 > \sqrt{\frac{A_{tot}}{\pi N_{AP}{}^*}} = D^*(N_{AP}). \tag{11.24}$$

When base station ranges substantially exceed the critical radius for full coverage, our system is likely to be interference limited and the analysis techniques from Chapter 5 apply. For a system capacity perspective this is the most interesting case. When operating a system close to its capacity it will be generating most revenue and it will be interference limited. Returning to expression (11.14) and combining it with (11.6) yields:

$$C_{sys} \approx c_2 \frac{R_{tot}}{\eta \bar{S} W_{sys}} + c_3 W_{sys} = c_2 \frac{\omega A_{tot} R_{user}}{\eta \bar{S} W_{sys}} + c_3 W_{sys}, \tag{11.25}$$

where we note that the expected number of users $N_{user} = \omega A_{tot}$ have to share the data rate R_{tot}, giving them an average rate of R_{user}. Examining this expression somewhat more closely, we see the three different ways to lower the system cost. Assuming the price of spectrum c_3 is constant we have:

I. Improve the efficiency of the modulation and RRM system, i.e. increase η.
II. Reduce the coverage area A_{tot}. The required data rate is only provided in parts of the area as illustrated by Figure 11.5.
III. Buy more spectrum—as long as the spectrum cost is below $C_3{}^*$ (Equation (11.16)).
IV. Reduce the cost per base station.

Improving the radio transmission technology, as in I, is a proven path. The achievable data rates in standardized mobile systems have increased dramatically over the last ten

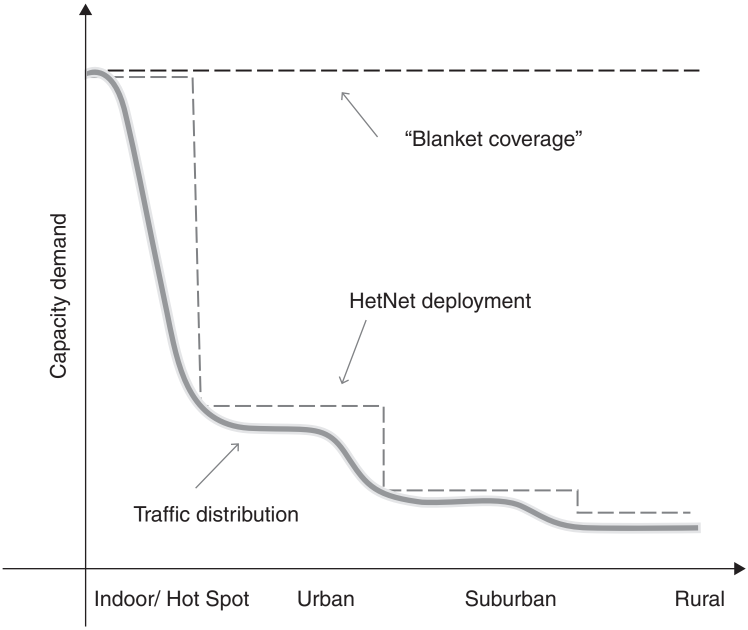


Figure 11.5 Capacity demand and deployment strategies for non-homogenous traffic.

years. In the 3GPP domain we have seen 3G/UMTS systems being enhanced by HSPA that are now, by many operators, being replaced or complemented by LTE systems. Driven by tremendous advances in signal processing capabilities, peak rates have risen from a few 100 kbits/sec to approach 100 Mbps in the latest LTE releases. Higher peak rates have been partially driven by better signal processing techniques that have benefited from Moore's law. This may have led to the common misunderstanding that future capacity/cost problems can be solved by simply replacing current equipment with new models with higher peak rates at lower cost. As we have seen in Chapter 5 already, it is not the peak rate that is critical, it is the efficiency η that tells us how close to the Shannon limit we may operate and how effective we are in mitigating interference. Novel, ingenious DSP schemes use multiple antenna systems, base station cooperation (CoMP) and interference control (inter-cell interference coordination, ICIC) techniques to increase efficiency. These techniques are often facilitated through centralized radio resource management, where processed, digital baseband waveforms are sent to simple radio heads (up-converters, power amplifiers and antennas) by means of optical fibers. The ultimate objective of these schemes is to effectively eliminate the co-channel interference caused by frequency reuse. Ideally this could result in a two- to seven-fold improvement in capacity. The gain, of course, will vary due to the amount of actual interference. In open areas with little environmental protection from interference, these schemes are likely to excel. In indoor environments with plenty of walls to shield interference, the benefits are likely to be limited [D. H. Kang et al., 2012]. Recent results from 3GPP standardization [WG1 3GPP TSG-RAN, 2010] show that such schemes seem to be very sensitive to estimation errors and delays in the channel state information that has to be passed around between the base stations. Under realistic conditions a factor of 2–3 improvements seems to be a plausible maximum benefit. An improvement, no doubt, but a rather modest one, for introducing a very high complexity.

Reducing the area coverage has also been a proven path. High capacities are only provided in certain areas, so-called hotspots, where there are many users, whereas more sparsely populated areas can get by with less capacity. Making sure to provide capacity where it is actually needed is the strategy. This is illustrated in Figure 11.5, which shows the strong variations of capacity demand for various locales, ranging from small areas indoors and in hotspots where the user densities are very high all the way to large rural areas with low capacity demands. Providing blanket coverage, i.e., a capacity corresponding to the peak demand everywhere (dark dashed line) would require a very large number of base stations. The light dashed line corresponds to heterogeneous network (HetNet) deployment where the service area is subdivided according to the demand. Very high capacities can be provided in small areas with very high demands at moderate costs since the corresponding $A_{service}$ in Equation (11.4) is small. In larger and larger areas less and less capacity is required. HetNet deployment tailors the provided capacity to the traffic demand, and can include several cellular tiers meshed together. As the demand is expected to increase most where the majority of users are, i.e. indoors, this is also where most of the most deployment is going to occur as we meet the demand.

Well-designed HetNets not only minimize the number of base stations, they also significantly reduce the cost per base station (item IV above) and the required energy consumption in the very small cells. Equation (11.26) shows the base station cost broken down to its major components:

$$C_{BS} = C_{site} + C_{backhaul} + C_{equipment} + C_{deployment} + C_{mast}. \tag{11.26}$$

For conventional outdoor, wide-area systems, all five components are in play. Site costs are high due to high masts, housing for the equipment, and roads for maintenance. Backhaul is also expensive since a macro site often requires the new installation of a dedicated fiber connection. Finally, equipment is high-power, industry grade equipment that often requires cooling; all this leads to a high price tag (10,000s of dollars). Moreover the deployment, planning and installation require skilled personnel and maintenance. Due to costs there is only limited redundancy built in for the wide-area coverage.

Figure 11.6 provides an example of this. It shows a map of an area where the traffic density varies as indicated by the different degrees of shading (the darker, the higher the traffic density). The task here was to deploy 3G (WCDMA) macro cells and WiFi access points to meet a certain user requirement at the lowest possible cost. In this case, the requirement was that 95% of the users should have access to a certain data rate. Figure 11.6 illustrates the solution to this optimization problem—the squares indicate the three macro cell (3G) towers that cover the area and grey dots the WiFi access points that provide high capacity at dense user locations.

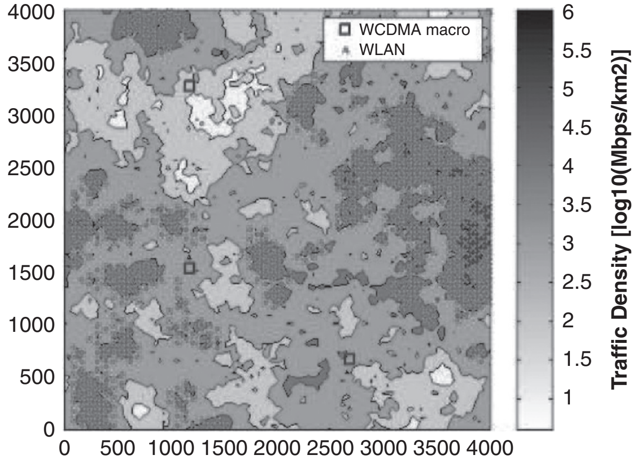


Figure 11.6 Example of practical HetNet deployment optimization—wide-area 3G + local WiFi [K. Johansson, 2007].

Indoor systems are deployed on the premises, usually by the facility owners, resulting in negligible site costs. The equipment is consumer grade and low cost. If the capacity is too low, a few more access points can easily be deployed. Coverage is not a problem and when a base station breaks this will only cause a minor degradation in capacity. The equipment can be replaced by the electrician at some later time, resulting in much lower maintenance costs. If the network is deployed indoors, where in most cases there is already a fiber or Ethernet outlet, the backhaul cost also vanishes—provided that the base stations can use standard IP access. If the network has the capability to self-organize, the dominant, remaining cost is that of the physical deployment of the base stations.

The analogy with lighting is striking—outdoor, public lighting is provided by large, high-power floodlights on high towers, whereas indoor lighting is provided by an abundance of simple lamps. This analogy also demonstrates another point—no one would dream of providing indoor lighting using street floodlights. Most of the energy would be wasted and absorbed in walls and the indoor illumination results would be poor. This is, on the other hand, exactly the solution we still see in mobile data operation today causing poor energy efficiency, high transmission powers and low indoor data rates.

The reason for this is not technical. Instead, we have a clash of business models [J. Markendahl et al., 2009; J. Markendahl and M. Östen, 2010]. Public outdoor access is provided by mobile operators as a subscription-based service, whereas indoor WiFi systems are provided as part of the services when renting an office, similar to electricity or ventilation. The public operators cannot hope to deploy their own systems in every building, but still users want seamless transitions from the outdoor public system to the indoor private system. A key research question is thus to find efficient techniques and business models for sharing the indoor systems. Since a majority of the traffic growth will be carried by indoor, low-cost base stations, this will certainly have an impact on the manufacturing industry. WiFi is certainly not the solution to all local access problems. WiFi may work in many homes, but in very high-density deployments there are significant capacity shortcomings with the simple CSMA/CA access protocol. Regardless of the technically advanced content of a future indoor base station, it will still eventually be a \$100 box that someone puts on the wall.

Exercises

11.1 In a cellular system the path gain from an access port to a terminal follows the expression $G(r) = C - 35\log_{10}(r)$ (dB). The system is designed to provide digital wireless access at a minimum data rate of 1 Mbps. At full power at the design range, 5 km, where the system achieves full coverage, the signal to noise ratio, E_b/N_0, is 5 dB. A single channel of 5 MHz bandwidth is used. Assume the system to be noise limited.

a) Determine the area coverage (in percent) at 2 Mbps, 5 Mbps and 10 Mbps minimum data rate.
b) How many more access ports are required to achieve full coverage for the data rates in (a) if the transmit power is constant?

11.2 Consider the system in Example 11.1 for 2 Mbps user data rate and full coverage.

a) How much lower would the infrastructure cost be (using cost model I) if additional spectrum for a 5 MHz channel was acquired?
b) Repeat the calculation for the 10 Mbps case.
c) If the cost per base station is 100K euro, determine the engineering value of the spectrum per cell.

11.3 In the system in Exercise 11.1, what if, instead of increasing the number of base stations to achieve full coverage, a higher transmit power is used? Compare the power for the 5 and 10 Mbps systems to the 1 Mbps system.

References

K. Johansson 2007. *Cost Effective Deployment Strategies for Heterogeneous Wireless Networks*. Doctoral thesis, KTH Royal Institute of Technology.

D. H. Kang, K. W. Sung and J. Zander. 2012 (Dec. 3–7). Is multicell interference coordination worthwhile in indoor wireless broadband systems? In: *Proc. IEEE Global Communications Conference (GLOBECOM) 2012*.

J. Markendahl, M. Östen, J. Werding and B. G. Mölleryd. 2009. Business innovation strategies to reduce the revenue gap for wireless broadband services. *Communications & Strategies*, 72, 3rd quarter, 35.

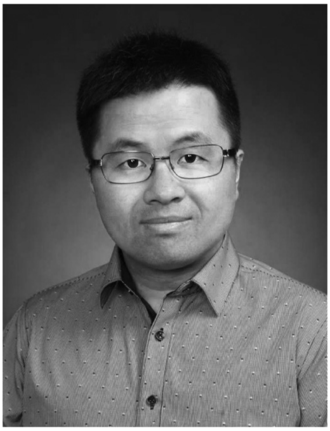
J. Markendahl and M. Östen. 2010 (Sept. 26–30). A comparative study of deployment options, capacity and cost structure for macrocellular and femtocell networks. Pages 145–150 of: *Proc. IEEE 21st International Symposium on Personal, Indoor and Mobile Radio Communications Workshops (PIMRC Workshops)*.

W. Sharkey 1982. *The Theory of Natural Monopoly*. Cambridge: Cambridge University Press.

WG1 3GPP TSG-RAN. 2010 (Feb.). *Performance Evaluation of Intra-Site DL CoMP*. R1-100855.

J. Zander and P. Mähönen. 2013. Riding the data tsunami in the cloud: myths and challenges in future wireless access. *IEEE Communications Magazine*, 51(3), 145–151.

About the authors



GUOWANG MIAO received B.S. and M.S. degrees from Tsinghua University and M.S. and Ph.D. degrees from Georgia Institute of Technology, Atlanta, GA, USA. He once worked at Intel Labs as a research engineer and at Samsung Research America as a senior standards engineer and 3GPP LTE-A delegate. In 2011, he won an Individual Gold Award from Samsung Research America for his contributions to LTE-A standardization. He joined KTH Royal Institute of Technology in February 2012 as an assistant professor; from February 2015, he has been a tenured associate professor in the same institution. His research interest is in the design and optimization of mobile communications and networking and he is well known for his contributions in energy-efficient communications. In addition to this book, he is the lead author of *Energy and Spectrum Efficient Wireless Network Design*, published by Cambridge University Press. He has published more than 60 research papers in premier journals or conferences. He has had several patents granted and many more filed. Several of his patents have been adopted in 4G standards. He has been a technical program committee member for many international conferences and been on the editorial board of several international journals. He was an exemplary reviewer for *IEEE Communications Letters* in 2011.

JENS ZANDER received his Ph.D. in electrical engineering from Linköping University, Sweden, in 1985. Between 1985 and 1990 he was co-founder and vice-president of SECTRA, a high-tech company in communication and security nowadays on the Swedish Stock Exchange. In 1989 he was appointed (full) professor and head of the Radio Communication Systems Laboratory at the Royal Institute of Technology, Stockholm, Sweden. He is co-founder of the Center for Wireless Systems (Wireless@KTH) at the Royal Institute of Technology, Stockholm and is currently Scientific Director of that center. He has been on the board of more than ten companies ranging from high-tech start-ups and venture capital firms to Teracom, the Swedish broadcasting operator. He is currently on the board of directors of the National Post and Telecom Agency (PTS). Dr Zander is a member of the Swedish Academy of Engineering Sciences (IVA) and a senior research advisor to the Swedish Defence Institute (FOI). His current research interests include architectures, resource and spectrum management regimes in wireless systems, as well as economic models for future wireless infrastructures.

KI WON SUNG is a docent researcher in the Communication Systems Department at KTH Royal Institute of Technology. He is also affiliated with the KTH Center for Wireless Systems (Wireless@KTH). He received a B.S. degree in industrial management, and M.S. and Ph.D. degrees in industrial engineering from Korea

Advanced Institute of Science and Technology (KAIST) in 1998, 2000 and 2005 respectively. From 2005 to 2007 he was a senior engineer at Samsung Electronics, Korea, where he participated in the development and commercialization of a mobile WiMAX system. In 2008 he was a visiting researcher at the Institute for Digital Communications, University of Edinburgh, Scotland. He joined KTH in 2009. He served as an assistant project coordinator of the European FP7 project QUASAR. He also served as a track chair for CROWNCOM 2012 and a TPC member for several international conferences. His research interests include dynamic spectrum access, energy-efficient wireless networks, cost-effective deployment and operation, and future wireless architecture.

SLIMANE BEN SLIMANE received his B.Sc. degree in electrical engineering from the University of Quebec in Trois-Rivières, Quebec, Canada in 1985, his M.Sc. degree from Concordia University, Montreal, Canada, in 1988, and his Ph.D. degree also from Concordia University in 1993. In October 1995, he joined the Department of Signals, Sensors, and Systems at the Royal Institute Technology as an assistant professor. Since then he has been involved in teaching modern radio communications and carrying out research projects. He is presently an associate professor in the area of radio communication. His research interest is in the area of wireless communications with special emphasis on digital communication techniques for fading channels, channel coding, access methods, cooperative communications, energy efficiency and cognitive radio.

Index

achievable SINR, 148
ALOHA, 44
  equilibrium, 49
  max throughput, 47
  pure ALOHA, 45
  slotted ALOHA, 46
association, 199, 221
  cell breathing, 222
  cell range extension, 223
  load balancing, 221
  network selection, 200
average power consumption
  slotted ALOHA, 249

BER, 30
best effort, 23
bit-map protocol, 58
blocking probability, 109
BS density, 253
  uplink, 72

capacity region
  downlink, 74
  duality, 75
  scheduling, 76
  two-user uplink, 73
cellular networks
  energy efficiency, 250
cellular systems
  best-effort data services, 110
  cluster size, 102
  directional antennas, 120
  orthognal multiple access, 95
  outage-based capacity, 114
  reuse distance, 100
  static channel allocation, 97
  traffic-based capacity, 106
channel borrowing, 185
CoMP, 277
  coordinated beamforming, 278
  coordinated scheduling, 278
  joint transmission, 278

contention based, 28
contention free, 28, 32
cost
  base station, 295
  subscriber, 288
cost model, 287, 291
CRA, 55
CSMA, 50
  1-persistent CSMA, 50
  CSMA/CA, 52, 53
  CSMA/CD, 52
  detection delay, 51
  non-persistent CSMA, 50
  p-persistent CSMA, 51

DCH, 62
delay, 29
directional antennas, 120
DS-CDMA systems, 122
  downlink capacity, 133
  interference factor, 218
  multi-service CDMA, 134
  sectorization, 129
  traffic-based capacity, 131
  uplink multi-cell capacity, 127
  uplink single-cell capacity, 126
  voice activity monitoring, 129

economy
  access, 286
EE and SE tradeoff
  interference, 244
energy consumption, 228
energy efficiency, 241
  channel gain, 233
  data rate
    circuit power, 233
  energy consumption, 229
  MIMO, 236
  OFDM, 236
  signal bandwidth, 233

energy-efficient
  network deployment, 250
energy-efficient MAC, 245
energy-efficient power allocation, 238
energy-efficient scheduling, 234
energy-efficient transmission, 381
  close form, 239
  frequency selective channels, 235
    necessary and sufficient condition, 237
  no circuit power, 230
  optimal data rate, 233
energy minimization, 229
  energy efficiency maximization, 229
engineering value, 291
equilibrium
  interference, 156
expected throughput, 77
exposed node problem, 53

fading margin, 204, 216
FDMA, 32
frequency hopping, 186

guard channels, 209

handover, 199
  frequency of handover, 203
  handover failure probability, 202
  hard handover, 201
  horizontal handover, 200
  MAHO, 201, 208
  soft handover, 201, 213
  vertical handover, 200
heterogeneous networks, 192, 223
  cell range extension, 223
  closed access, 193
  closed subscriber group, 193
  energy efficient, 254
  open access, 193
hidden node problem, 52
hybrid access, 38
hysteresis, 206, 208

IEEE 802.11
  NAV, 61
interference management
  interference avoidance, 178, 181
  interference cancellation, 178, 189
  interference randomization, 178, 186

license, 284
link gain matrix, 21
location management, 199
long distance network, 286
LTE
  contention based, 270
  contention free, 272
  diversity, 265
  downlink, 258
  frame structure, 66, 262
  interference management, 272
  multiple antennas, 262
  OFDMA, 258
  random access, 270
  SC-FDMA, 267
  soft frequency reuse, 272
  spatial multiplexing, 267
  uplink, 267

MAC
  cellular networks, 62
  IEEE 802.11, 60
multi-cell scheduling, 185
multiuser capacity, 72

network coupling factor, 242
network energy consumption, 251
network throughput, 47
noise, 4

OFDM, 258, 259
OFDMA, 36
  LTE, 258, 261
operation mode, 245

partial frequency reuse, 279
path loss, 3
PER, 30
  BER, 30
poll, 41
power allocation
  water-filling
    energy efficient, 237
power consumption, 245
  MAC, 246
  MIMO, 236
  OFDM, 236
power control
  achievable region, 159
  centralized, 149
  close loop, 170
  constrained, 157

power control (*cont.*)
  data service, 159
  distributed, 153, 162
    convergence, 154
  DS-CDMA, 166
  energy efficient
    conservative, 244
  open loop, 168
  transmitter removal, 152
proportional fairness, 79

QoS, 68
  UMTS, 17

RACH, 62
real-time service, 22
reservation ALOHA, 58
reservation-based protocol, 57
resource allocation
  dynamic, 25
  static, 24
reuse distance, 99
reuse partitioning, 181, 272
roaming, 199
RTS/CTS, 54, 61

SC-FDMA
  LTE, 269
scheduling, 67
  fairness and efficiency, 83
  max throughput, 76
    OFDMA, 88
  max utility, 81
    alpha fair, 83
  max–min, 80
  OFDMA, 87
  proportional fair, 78
  queue, 70
  queuing model, 70
  round-robin, 76
SDMA, 37
SE
  upper bound, 243
service classes, 17
SIR balancing, 150
spectral efficiency, 241
spectrum, 4
spread spectrum techniques, 187
SSMA, 34
standard interference function, 154
successive interference cancelation, 72, 189
system throughput, 31

TDMA, 33
throughput, 31
time to trigger, 208
token ring, 41
total energy consumption, 231
tradeoff, 10
  energy and spectral efficiency
    interference, 241, 243
    interference-free, 241
  energy efficiency and spectral efficiency, 241
  SE and EE, 243
  spectral and energy
    single user, 241
  spectral and energy efficiency, 241
transmission energy, 230
transmit beamforming, 190

water-filling power allocation, 89
wireless channel
  scheduling, 70

zero forcing, 191