

APPENDIX 3

4G

LTE / LTE-Advanced for Mobile Broadband

Erik Dahlman
Stefan Parkvall
Johan Sköld



4G LTE/LTE-Advanced for Mobile Broadband

4G LTE/LTE-Advanced for Mobile Broadband

**Erik Dahlman, Stefan Parkvall, and
Johan Sköld**



AMSTERDAM • BOSTON • HEIDELBERG • LONDON • NEW YORK • OXFORD
PARIS • SAN DIEGO • SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO

Academic Press is an imprint of Elsevier



Academic Press is an imprint of Elsevier
The Boulevard, Langford Lane, Kidlington, Oxford, OX5 1GB, UK
30 Corporate Drive, Suite 400, Burlington, MA 01803, USA

First published 2011

Copyright © 2011 Erik Dahlman, Stefan Parkvall & Johan Sköld. Published by Elsevier Ltd. All rights reserved

The rights of Erik Dahlman, Stefan Parkvall & Johan Sköld to be identified as the authors of this work has been asserted in accordance with the Copyright, Designs and Patents Act 1988.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangement with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

British Library Cataloguing-in-Publication Data

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

Library of Congress Control Number: 2011921244

ISBN: 978-0-12-385489-6

For information on all Academic Press publications
visit our website at www.elsevierdirect.com

Typeset by MPS Limited, a Macmillan Company, Chennai, India
www.macmillansolutions.com

Printed and bound in the UK

11 12 13 14 10 9 8 7 6 5 4 3 2 1

Working together to grow
libraries in developing countries

www.elsevier.com | www.bookaid.org | www.sabre.org

ELSEVIER

BOOK AID
International

Sabre Foundation

Preface

During the past years, there has been a quickly rising interest in radio access technologies for providing mobile as well as nomadic and fixed services for voice, video, and data. The difference in design, implementation, and use between telecom and datacom technologies is also becoming more blurred. One example is cellular technologies from the telecom world being used for broadband data and wireless LAN from the datacom world being used for voice-over IP.

Today, the most widespread radio access technology for mobile communication is digital cellular, with the number of users passing 5 billion by 2010, which is more than half of the world's population. It has emerged from early deployments of an expensive voice service for a few car-borne users, to today's widespread use of mobile-communication devices that provide a range of mobile services and often include camera, MP3 player, and PDA functions. With this widespread use and increasing interest in mobile communication, a continuing evolution ahead is foreseen.

This book describes LTE, developed in 3GPP (*Third Generation Partnership Project*) and providing true 4G broadband mobile access, starting from the first version in release 8 and through the continuing evolution to release 10, the latest version of LTE. Release 10, also known as LTE-Advanced, is of particular interest as it is the major technology approved by the ITU as fulfilling the IMT-Advanced requirements. The description in this book is based on LTE release 10 and thus provides a complete description of the LTE-Advanced radio access from the bottom up.

Chapter 1 gives the background to LTE and its evolution, looking also at the different standards bodies and organizations involved in the process of defining 4G. It also gives a discussion of the reasons and driving forces behind the evolution.

Chapters 2–6 provide a deeper insight into some of the technologies that are part of LTE and its evolution. Because of its generic nature, these chapters can be used as a background not only for LTE as described in this book, but also for readers who want to understand the technology behind other systems, such as WCDMA/HSPA, WiMAX, and CDMA2000.

Chapters 7–17 constitute the main part of the book. As a start, an introductory technical overview of LTE is given, where the most important technology components are introduced based on the generic technologies described in previous chapters. The following chapters provide a detailed description of the protocol structure, the downlink and uplink transmission schemes, and the associated mechanisms for scheduling, retransmission and interference handling. Broadcast operation and relaying are also described. This is followed by a discussion of the spectrum flexibility and the associated requirements from an RF perspective.

Finally, in Chapters 18–20, an assessment is made on LTE. Through an overview of similar technologies developed in other standards bodies, it will be clear that the technologies adopted for the evolution in 3GPP are implemented in many other systems as well. Finally, looking into the future, it will be seen that the evolution does not stop with LTE-Advanced but that new features are continuously added to LTE in order to meet future requirements.

Acknowledgements

We thank all our colleagues at Ericsson for assisting in this project by helping with contributions to the book, giving suggestions and comments on the contents, and taking part in the huge team effort of developing LTE.

The standardization process involves people from all parts of the world, and we acknowledge the efforts of our colleagues in the wireless industry in general and in 3GPP RAN in particular. Without their work and contributions to the standardization, this book would not have been possible.

Finally, we are immensely grateful to our families for bearing with us and supporting us during the long process of writing this book.

Abbreviations and Acronyms

3GPP	Third Generation Partnership Project
3GPP2	Third Generation Partnership Project 2
ACIR	Adjacent Channel Interference Ratio
ACK	Acknowledgement (in ARQ protocols)
ACLR	Adjacent Channel Leakage Ratio
ACS	Adjacent Channel Selectivity
AM	Acknowledged Mode (RLC configuration)
AMC	Adaptive Modulation and Coding
A-MPR	Additional Maximum Power Reduction
AMPS	Advanced Mobile Phone System
AQPSK	Adaptive QPSK
ARI	Acknowledgement Resource Indicator
ARIB	Association of Radio Industries and Businesses
ARQ	Automatic Repeat-reQuest
AS	Access Stratum
ATIS	Alliance for Telecommunications Industry Solutions
AWGN	Additive White Gaussian Noise
BC	Band Category
BCCH	Broadcast Control Channel
BCH	Broadcast Channel
BER	Bit-Error Rate
BLER	Block-Error Rate
BM-SC	Broadcast Multicast Service Center
BPSK	Binary Phase-Shift Keying
BS	Base Station
BSC	Base Station Controller
BTS	Base Transceiver Station
CA	Carrier Aggregation
CC	Convolutional Code (in the context of coding), or Component Carrier (in the context of carrier aggregation)
CCCH	Common Control Channel
CCE	Control Channel Element
CCSA	China Communications Standards Association
CDD	Cyclic-Delay Diversity
CDF	Cumulative Density Function
CDM	Code-Division Multiplexing
CDMA	Code-Division Multiple Access

CEPT	European Conference of Postal and Telecommunications Administrations
CN	Core Network
CoMP	Coordinated Multi-Point transmission/reception
CP	Cyclic Prefix
CPC	Continuous Packet Connectivity
CQI	Channel-Quality Indicator
C-RAN	Centralized RAN
CRC	Cyclic Redundancy Check
C-RNTI	Cell Radio-Network Temporary Identifier
CRS	Cell-specific Reference Signal
CS	Circuit Switched (or Cyclic Shift)
CS	Capability Set (for MSR base stations)
CSA	Common Subframe Allocation
CSG	Closed Subscriber Group
CSI	Channel-State Information
CSI-RS	CSI reference signals
CW	Continuous Wave
DAI	Downlink Assignment Index
DCCH	Dedicated Control Channel
DCH	Dedicated Channel
DCI	Downlink Control Information
DFE	Decision-Feedback Equalization
DFT	Discrete Fourier Transform
DFTS-OFDM	DFT-Spread OFDM (DFT-precoded OFDM, see also SC-FDMA)
DL	Downlink
DL-SCH	Downlink Shared Channel
DM-RS	Demodulation Reference Signal
DRX	Discontinuous Reception
DTCH	Dedicated Traffic Channel
DTX	Discontinuous Transmission
DwPTS	The downlink part of the special subframe (for TDD operation).
EDGE	Enhanced Data rates for GSM Evolution, Enhanced Data rates for Global Evolution
EGPRS	Enhanced GPRS
eNB	eNodeB
eNodeB	E-UTRAN NodeB
EPC	Evolved Packet Core
EPS	Evolved Packet System
ETSI	European Telecommunications Standards Institute
E-UTRA	Evolved UTRA
E-UTRAN	Evolved UTRAN
EV-DO	Evolution-Data Only (of CDMA2000 1x)
EV-DV	Evolution-Data and Voice (of CDMA2000 1x)
EVM	Error Vector Magnitude

FACH	Forward Access Channel
FCC	Federal Communications Commission
FDD	Frequency Division Duplex
FDM	Frequency-Division Multiplex
FDMA	Frequency-Division Multiple Access
FEC	Forward Error Correction
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
FPLMTS	Future Public Land Mobile Telecommunications Systems
FRAMES	Future Radio Wideband Multiple Access Systems
FSTD	Frequency Switched Transmit Diversity
GERAN	GSM/EDGE Radio Access Network
GGSN	Gateway GPRS Support Node
GP	Guard Period (for TDD operation)
GPRS	General Packet Radio Services
GPS	Global Positioning System
GSM	Global System for Mobile communications
HARQ	Hybrid ARQ
HII	High-Interference Indicator
HLR	Home Location Register
HRPD	High Rate Packet Data
HSDPA	High-Speed Downlink Packet Access
HSPA	High-Speed Packet Access
HSS	Home Subscriber Server
HS-SCCH	High-Speed Shared Control Channel
ICIC	Inter-Cell Interference Coordination
ICS	In-Channel Selectivity
ICT	Information and Communication Technologies
IDFT	Inverse DFT
IEEE	Institute of Electrical and Electronics Engineers
IFDMA	Interleaved FDMA
IFFT	Inverse Fast Fourier Transform
IMT-2000	International Mobile Telecommunications 2000 (ITU's name for the family of 3G standards)
IMT-Advanced	International Mobile Telecommunications Advanced (ITU's name for the family of 4G standards)
IP	Internet Protocol
IR	Incremental Redundancy
IRC	Interference Rejection Combining
ITU	International Telecommunications Union
ITU-R	International Telecommunications Union-Radiocommunications Sector

J-TACS	Japanese Total Access Communication System
LAN	Local Area Network
LCID	Logical Channel Index
LDPC	Low-Density Parity Check Code
LTE	Long-Term Evolution
MAC	Medium Access Control
MAN	Metropolitan Area Network
MBMS	Multimedia Broadcast/Multicast Service
MBMS-GW	MBMS gateway
MBS	Multicast and Broadcast Service
MBSFN	Multicast-Broadcast Single Frequency Network
MC	Multi-Carrier
MCCH	MBMS Control Channel
MCE	MBMS Coordination Entity
MCH	Multicast Channel
MCS	Modulation and Coding Scheme
MDHO	Macro-Diversity Handover
MIB	Master Information Block
MIMO	Multiple-Input Multiple-Output
ML	Maximum Likelihood
MLSE	Maximum-Likelihood Sequence Estimation
MME	Mobility Management Entity
MMS	Multimedia Messaging Service
MMSE	Minimum Mean Square Error
MPR	Maximum Power Reduction
MRC	Maximum Ratio Combining
MSA	MCH Subframe Allocation
MSC	Mobile Switching Center
MSI	MCH Scheduling Information
MSP	MCH Scheduling Period
MSR	Multi-Standard Radio
MSS	Mobile Satellite Service
MTCH	MBMS Traffic Channel
MU-MIMO	Multi-User MIMO
MUX	Multiplexer or Multiplexing
NAK, NACK	Negative Acknowledgement (in ARQ protocols)
NAS	Non-Access Stratum (a functional layer between the core network and the terminal that supports signaling and user data transfer)
NDI	New-data indicator
NSPS	National Security and Public Safety
NMT	Nordisk MobilTelefon (Nordic Mobile Telephony)

NodeB	NodeB, a logical node handling transmission/reception in multiple cells. Commonly, but not necessarily, corresponding to a base station.
NS	Network Signaling
OCC	Orthogonal Cover Code
OFDM	Orthogonal Frequency-Division Multiplexing
OFDMA	Orthogonal Frequency-Division Multiple Access
OI	Overload Indicator
OOB	Out-Of-Band (emissions)
PAPR	Peak-to-Average Power Ratio
PAR	Peak-to-Average Ratio (same as PAPR)
PARC	Per-Antenna Rate Control
PBCH	Physical Broadcast Channel
PCCH	Paging Control Channel
PCFICH	Physical Control Format Indicator Channel
PCG	Project Coordination Group (in 3GPP)
PCH	Paging Channel
PCRF	Policy and Charging Rules Function
PCS	Personal Communications Systems
PDA	Personal Digital Assistant
PDC	Personal Digital Cellular
PDCCH	Physical Downlink Control Channel
PDCP	Packet Data Convergence Protocol
PDSCH	Physical Downlink Shared Channel
PDN	Packet Data Network
PDU	Protocol Data Unit
PF	Proportional Fair (a type of scheduler)
P-GW	Packet-Data Network Gateway (also PDN-GW)
PHICH	Physical Hybrid-ARQ Indicator Channel
PHS	Personal Handy-phone System
PHY	Physical layer
PMCH	Physical Multicast Channel
PMI	Precoding-Matrix Indicator
POTS	Plain Old Telephony Services
PRACH	Physical Random Access Channel
PRB	Physical Resource Block
P-RNTI	Paging RNTI
PS	Packet Switched
PSK	Phase Shift Keying
PSS	Primary Synchronization Signal
PSTN	Public Switched Telephone Networks
PUCCH	Physical Uplink Control Channel

PUSC	Partially Used Subcarriers (for WiMAX)
PUSCH	Physical Uplink Shared Channel
QAM	Quadrature Amplitude Modulation
QoS	Quality-of-Service
QPP	Quadrature Permutation Polynomial
QPSK	Quadrature Phase-Shift Keying
RAB	Radio Access Bearer
RACH	Random Access Channel
RAN	Radio Access Network
RA-RNTI	Random Access RNTI
RAT	Radio Access Technology
RB	Resource Block
RE	Resource Element
RF	Radio Frequency
RI	Rank Indicator
RIT	Radio Interface Technology
RLC	Radio Link Control
RNC	Radio Network Controller
RNTI	Radio-Network Temporary Identifier
RNTP	Relative Narrowband Transmit Power
ROHC	Robust Header Compression
R-PDCCH	Relay Physical Downlink Control Channel
RR	Round-Robin (a type of scheduler)
RRC	Radio Resource Control
RRM	Radio Resource Management
RS	Reference Symbol
RSPC	IMT-2000 radio interface specifications
RSRP	Reference Signal Received Power
RSRQ	Reference Signal Received Quality
RTP	Real Time Protocol
RTT	Round-Trip Time
RV	Redundancy Version
RX	Receiver
S1	The interface between eNodeB and the Evolved Packet Core.
S1-c	The control-plane part of S1
S1-u	The user-plane part of S1
SAE	System Architecture Evolution
SCM	Spatial Channel Model
SDMA	Spatial Division Multiple Access
SDO	Standards Developing Organization
SDU	Service Data Unit
SEM	Spectrum Emissions Mask

SF	Spreading Factor
SFBC	Space-Frequency Block Coding
SFN	Single-Frequency Network (in general, see also MBSFN) or System Frame Number (in 3GPP)
SFTD	Space-Frequency Time Diversity
SGSN	Serving GPRS Support Node
S-GW	Serving Gateway
SI	System Information message
SIB	System Information Block
SIC	Successive Interference Combining
SIM	Subscriber Identity Module
SINR	Signal-to-Interference-and-Noise Ratio
SIR	Signal-to-Interference Ratio
SI-RNTI	System Information RNTI
SMS	Short Message Service
SNR	Signal-to-Noise Ratio
SOHO	Soft Handover
SORTD	Spatial Orthogonal-Resource Transmit Diversity
SR	Scheduling Request
SRS	Sounding Reference Signal
SSS	Secondary Synchronization Signal
STBC	Space-Time Block Coding
STC	Space-Time Coding
STTD	Space-Time Transmit Diversity
SU-MIMO	Single-User MIMO
TACS	Total Access Communication System
TCP	Transmission Control Protocol
TC-RNTI	Temporary C-RNTI
TD-CDMA	Time-Division Code-Division Multiple Access
TDD	Time-Division Duplex
TDM	Time-Division Multiplexing
TDMA	Time-Division Multiple Access
TD-SCDMA	Time-Division-Synchronous Code-Division Multiple Access
TF	Transport Format
TIA	Telecommunications Industry Association
TM	Transparent Mode (RLC configuration)
TR	Technical Report
TS	Technical Specification
TSG	Technical Specification Group
TTA	Telecommunications Technology Association
TTC	Telecommunications Technology Committee
TTI	Transmission Time Interval
TX	Transmitter

UCI	Uplink Control Information
UE	User Equipment, the 3GPP name for the mobile terminal
UL	Uplink
UL-SCH	Uplink Shared Channel
UM	Unacknowledged Mode (RLC configuration)
UMB	Ultra Mobile Broadband
UMTS	Universal Mobile Telecommunications System
UpPTS	The uplink part of the special subframe (for TDD operation).
US-TDMA	US Time-Division Multiple Access standard
UTRA	Universal Terrestrial Radio Access
UTRAN	Universal Terrestrial Radio Access Network
VAMOS	Voice services over Adaptive Multi-user channels
VoIP	Voice-over-IP
VRB	Virtual Resource Block
WAN	Wide Area Network
WARC	World Administrative Radio Congress
WCDMA	Wideband Code-Division Multiple Access
WG	Working Group
WiMAX	Worldwide Interoperability for Microwave Access
WLAN	Wireless Local Area Network
WMAN	Wireless Metropolitan Area Network
WP5D	Working Party 5D
WRC	World Radiocommunication Conference
X2	The interface between eNodeBs.
ZC	Zadoff-Chu
ZF	Zero Forcing

Background of LTE

1

1.1 INTRODUCTION

Mobile communications has become an everyday commodity. In the last decades, it has evolved from being an expensive technology for a few selected individuals to today's ubiquitous systems used by a majority of the world's population. From the first experiments with radio communication by Guglielmo Marconi in the 1890s, the road to truly mobile radio communication has been quite long. To understand the complex mobile-communication systems of today, it is important to understand where they came from and how cellular systems have evolved. The task of developing mobile technologies has also changed, from being a national or regional concern, to becoming an increasingly complex task undertaken by global standards-developing organizations such as the *Third Generation Partnership Project* (3GPP) and involving thousands of people.

Mobile communication technologies are often divided into generations, with 1G being the analog mobile radio systems of the 1980s, 2G the first digital mobile systems, and 3G the first mobile systems handling broadband data. The *Long-Term Evolution* (LTE) is often called "4G", but many also claim that LTE release 10, also referred to as *LTE-Advanced*, is the true 4G evolution step, with the first release of LTE (release 8) then being labeled as "3.9G". This continuing race of increasing sequence numbers of mobile system generations is in fact just a matter of labels. What is important is the actual system capabilities and how they have evolved, which is the topic of this chapter.

In this context, it must first be pointed out that LTE and LTE-Advanced is the same technology, with the "Advanced" label primarily being added to highlight the relation between LTE release 10 (LTE-Advanced) and ITU/IMT-Advanced, as discussed later. This does not make LTE-Advanced a different system than LTE and it is not in any way the final evolution step to be taken for LTE. Another important aspect is that the work on developing LTE and LTE-Advanced is performed as a continuing task within 3GPP, the same forum that developed the first 3G system (WCDMA/HSPA).

This chapter describes the background for the development of the LTE system, in terms of events, activities, organizations and other factors that have played an important role. First, the technologies and mobile systems leading up to the starting point for 3G mobile systems will be discussed. Next, international activities in the ITU that were part of shaping 3G and the 3G evolution and the market and technology drivers behind LTE will be discussed. The final part of the chapter describes the standardization process that provided the detailed specification work leading to the LTE systems deployed and in operation today.

1.2 EVOLUTION OF MOBILE SYSTEMS BEFORE LTE

The US *Federal Communications Commission* (FCC) approved the first commercial car-borne telephony service in 1946, operated by AT&T. In 1947 AT&T also introduced the cellular concept of reusing radio frequencies, which became fundamental to all subsequent mobile-communication systems. Similar systems were operated by several monopoly telephone administrations and wire-line operators during the 1950s and 1960s, using bulky and power-hungry equipment and providing car-borne services for a very limited number of users.

The big uptake of subscribers and usage came when mobile communication became an international concern involving several interested parties, in the beginning mainly the operators. The first international mobile communication systems were started in the early 1980s; the best-known ones are NMT that was started up in the Nordic countries, AMPS in the USA, TACS in Europe, and J-TACS in Japan. Equipment was still bulky, mainly car-borne, and voice quality was often inconsistent, with “cross-talk” between users being a common problem. With NMT came the concept of “roaming”, giving a service also for users traveling outside the area of their “home” operator. This also gave a larger market for mobile phones, attracting more companies into the mobile-communication business.

The analog first-generation cellular systems supported “plain old telephony services” (POTS) – that is, voice with some related supplementary services. With the advent of digital communication during the 1980s, the opportunity to develop a second generation of mobile-communication standards and systems, based on digital technology, surfaced. With digital technology came an opportunity to increase the capacity of the systems, to give a more consistent quality of the service, and to develop much more attractive and truly mobile devices.

In Europe, the GSM (originally *Groupe Spécial Mobile*, later *Global System for Mobile communications*) project to develop a pan-European mobile-telephony system was initiated in the mid 1980s by the telecommunication administrations in CEPT¹ and later continued within the new *European Telecommunication Standards Institute* (ETSI). The GSM standard was based on *Time-Division Multiple Access* (TDMA), as were the US-TDMA standard and the Japanese PDC standard that were introduced in the same time frame. A somewhat later development of a *Code-Division Multiple Access* (CDMA) standard called IS-95 was completed in the USA in 1993.

All these standards were “narrowband” in the sense that they targeted “low-bandwidth” services such as voice. With the second-generation digital mobile communications came also the opportunity to provide data services over the mobile-communication networks. The primary data services introduced in 2G were text messaging (*Short Message Services*, SMS) and circuit-switched data services enabling e-mail and other data applications, initially at a modest peak data rate of 9.6 kbit/s. Higher data rates were introduced later in evolved 2G systems by assigning multiple time slots to a user and through modified coding schemes.

Packet data over cellular systems became a reality during the second half of the 1990s, with *General Packet Radio Services* (GPRS) introduced in GSM and packet data also added to other cellular technologies such as the Japanese PDC standard. These technologies are often referred to as 2.5G. The success of the wireless data service iMode in Japan, which included a complete “ecosystem”

¹ The European Conference of Postal and Telecommunications Administrations (CEPT) consists of the telecom administrations from 48 countries.

for service delivery, charging etc., gave a very clear indication of the potential for applications over packet data in mobile systems, in spite of the fairly low data rates supported at the time.

With the advent of 3G and the higher-bandwidth radio interface of UTRA (*Universal Terrestrial Radio Access*) came possibilities for a range of new services that were only hinted at with 2G and 2.5G. The 3G radio access development is today handled in 3GPP. However, the initial steps for 3G were taken in the early 1990s, long before 3GPP was formed.

What also set the stage for 3G was the internationalization of cellular standards. GSM was a pan-European project, but it quickly attracted worldwide interest when the GSM standard was deployed in a number of countries outside Europe. A global standard gains in economy of scale, since the market for products becomes larger. This has driven a much tighter international cooperation around 3G cellular technologies than for the earlier generations.

1.2.1 The First 3G Standardization

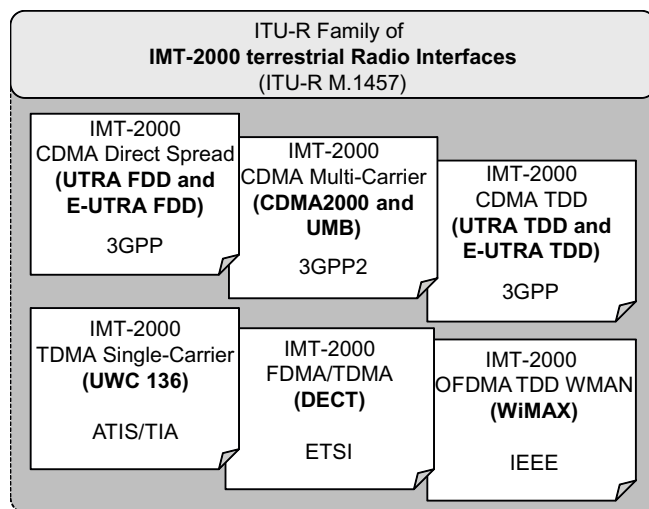
Work on a third-generation mobile communication started in ITU (*International Telecommunication Union*) in the 1980s, first under the label *Future Public Land Mobile Telecommunications Systems* (FPLMTS), later changed to IMT-2000 [1]. The World Administrative Radio Congress WARC-92 identified 230 MHz of spectrum for IMT-2000 on a worldwide basis. Of these 230 MHz, 2×60 MHz was identified as paired spectrum for FDD (*Frequency-Division Duplex*) and 35 MHz as unpaired spectrum for TDD (*Time-Division Duplex*), both for terrestrial use. Some spectrum was also set aside for satellite services. With that, the stage was set to specify IMT-2000.

In parallel with the widespread deployment and evolution of 2G mobile-communication systems during the 1990s, substantial efforts were put into 3G research activities worldwide. In Europe, a number of partially EU-funded projects resulted in a multiple access concept that included a Wideband CDMA component that was input to ETSI in 1996. In Japan, the *Association of Radio Industries and Businesses* (ARIB) was at the same time defining a 3G wireless communication technology based on Wideband CDMA and also in the USA a Wideband CDMA concept called WIMS was developed within the T1.P1² committee. South Korea also started work on Wideband CDMA at this time.

When the standardization activities for 3G started in ETSI in 1996, there were WCDMA concepts proposed both from a European research project (FRAMES) and from the ARIB standardization in Japan. The Wideband CDMA proposals from Europe and Japan were merged and came out as part of the winning concept in early 1998 in the European work on Universal Mobile Telecommunication Services (UMTS), which was the European name for 3G. Standardization of WCDMA continued in parallel in several standards groups until the end of 1998, when the *Third Generation Partnership Project* (3GPP) was formed by standards-developing organizations from all regions of the world. This solved the problem of trying to maintain parallel development of aligned specifications in multiple regions. The present organizational partners of 3GPP are ARIB (Japan), CCSA (China), ETSI (Europe), ATIS (USA), TTA (South Korea), and TTC (Japan).

At this time, when the standardization bodies were ready to put the details into the 3GPP specifications, work on 3G mobile systems had already been ongoing for some time in the international arena within the ITU-R. That work was influenced by and also provided a broader international framework for the standardization work in 3GPP.

²The T1.P1 committee was part of T1, which presently has joined the ATIS standardization organization.

**FIGURE 1.1**

The definition of IMT-2000 in ITU-R.

1.3 ITU ACTIVITIES

1.3.1 IMT-2000 and IMT-Advanced

ITU-R *Working Party 5D* (WP5D) has the responsibility for IMT systems, which is the umbrella name for 3G (IMT-2000) and 4G (IMT-Advanced). WP5D does not write technical specifications for IMT, but has kept the role of defining IMT in cooperation with the regional standardization bodies and to maintain a set of recommendations for IMT-2000 and IMT-Advanced.

The main IMT-2000 recommendation is ITU-R M.1457 [2], which identifies the IMT-2000 *radio interface specifications* (RSPC). The recommendation contains a “family” of radio interfaces, all included on an equal basis. The family of six terrestrial radio interfaces is illustrated in [Figure 1.1](#), which also shows the *Standards Developing Organizations* (SDO) or Partnership Projects that produce the specifications. In addition, there are several IMT-2000 satellite radio interfaces defined, not illustrated in [Figure 1.1](#).

For each radio interface, M.1457 contains an overview of the radio interface, followed by a list of references to the detailed specifications. The actual specifications are maintained by the individual SDOs and M.1457 provides references to the specifications maintained by each SDO.

With the continuing development of the IMT-2000 radio interfaces, including the evolution of UTRA to Evolved UTRA, the ITU recommendations also need to be updated. ITU-R WP5D continuously revises recommendation M.1457 and at the time of writing it is in its ninth version. Input to the updates is provided by the SDOs and Partnership Projects writing the standards. In the latest revision of ITU-R M.1457, LTE (or E-UTRA) is included in the family through the 3GPP family members for UTRA FDD and TDD, as shown in the figure.

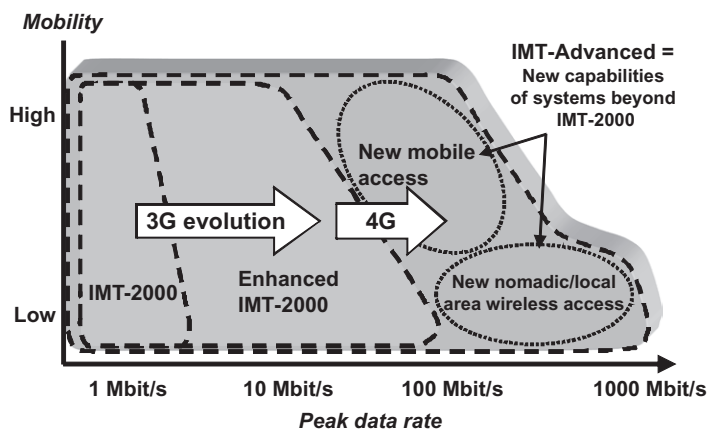


FIGURE 1.2

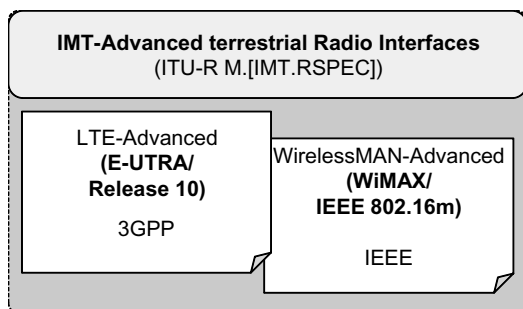
Illustration of capabilities of IMT-2000 and IMT-Advanced, based on the framework described in ITU -R Recommendation M.1645 [4].

In addition to maintaining the IMT-2000 specifications, a main activity in ITU-R WP5D is the work on systems beyond IMT-2000, now called IMT-Advanced. The term IMT-Advanced is used for systems that include new radio interfaces supporting the new capabilities of systems beyond IMT-2000, as demonstrated with the “van diagram” in Figure 1.2. The step into IMT-Advanced capabilities is seen by ITU-R as the step into 4G, the next generation of mobile technologies after 3G.

The process for defining IMT-Advanced was set by ITU-R WP5D [3] and was quite similar to the process used in developing the IMT-2000 recommendations. ITU-R first concluded studies for IMT-Advanced of services and technologies, market forecasts, principles for standardization, estimation of spectrum needs, and identification of candidate frequency bands [4]. Evaluation criteria were agreed, where proposed technologies were to be evaluated according to a set of minimum technical requirements. All ITU members and other organizations were then invited to the process through a circular letter [5] in March 2008. After submission of six candidate technologies in 2009, an evaluation was performed in cooperation with external bodies such as standards-developing organizations, industry forums, and national groups.

An evolution of LTE as developed by 3GPP was submitted as one candidate to the ITU-R evaluation. While actually being a new release (release 10) of the LTE system and thus an integral part of the continuing LTE development, the candidate was named LTE-Advanced for the purpose of ITU submission. 3GPP also set up its own set of technical requirements for LTE-Advanced, with the ITU-R requirements as a basis. The specifics of LTE-Advanced will be described in more detail as part of the description of LTE later in this book. The performance evaluation of LTE-Advanced for the ITU-R submission is described further in Chapter 18.

The target of the process was always harmonization of the candidates through consensus building. ITU-R determined in October 2010 that two technologies will be included in the first release of IMT-Advanced, those two being LTE release 10 (“LTE-Advanced”) and WirelessMAN-Advanced [6] based on the IEEE 802.16m specification. The two can be viewed as the “family” of IMT-Advanced

**FIGURE 1.3**

Radio interface technologies for IMT-Advanced.

technologies as shown in [Figure 1.3](#). The main IMT-Advanced recommendation, identifying the IMT-Advanced radio interface specifications, is presently named ITU-R.[IMT.RSPEC] [7] and will be completed during 2011. As for the corresponding IMT-2000 specification, it will contain an overview of each radio interface, followed by a list of references to the detailed specifications.

1.3.2 Spectrum for IMT Systems

Another major activity within ITU-R concerning IMT-Advanced has been to identify globally available spectrum, suitable for IMT systems. The spectrum work has involved sharing studies between IMT and other technologies in those bands. Adequate spectrum availability and globally harmonized spectrum are identified as essential for IMT-Advanced.

Spectrum for 3G was first identified at the *World Administrative Radio Congress* WARC-92, where 230 MHz was identified as intended for use by national administrations that want to implement IMT-2000. The so-called IMT-2000 “core band” at 2 GHz is in this frequency range and was the first band where 3G systems were deployed.

Additional spectrum was identified for IMT-2000 at later World Radio communication conferences. WRC-2000 identified the existing 2G bands at 800/900 MHz and 1800/1900 MHz plus an additional 190 MHz of spectrum at 2.6 GHz, all for IMT-2000. As additional spectrum for IMT-2000, WRC’07 identified a band at 450 MHz, the so-called “digital dividend” at 698–806 MHz, plus an additional 300 MHz of spectrum at higher frequencies. The applicability of these new bands varies on a regional and national basis.

The worldwide frequency arrangements for IMT-2000 are outlined in ITU-R recommendation M.1036 [8], which is presently being updated with the arrangements for the most recent frequency bands added at WRC’07. The recommendation outlines the regional variations in how the bands are implemented and also identifies which parts of the spectrum are paired and which are unpaired. For the paired spectrum, the bands for uplink (mobile transmit) and downlink (base-station transmit) are identified for *Frequency-Division Duplex* (FDD) operation. The unpaired bands can, for example, be used for *Time-Division Duplex* (TDD) operation. Note that the band that is most globally deployed for 3G is still 2 GHz.

1.4 DRIVERS FOR LTE

The evolution of 3G systems into 4G is driven by the creation and development of new services for mobile devices, and is enabled by advancement of the technology available for mobile systems. There has also been an evolution of the environment in which mobile systems are deployed and operated, in terms of competition between mobile operators, challenges from other mobile technologies, and new regulation of spectrum use and market aspects of mobile systems.

The rapid evolution of the technology used in telecommunication systems, consumer electronics, and specifically mobile devices has been remarkable in the last 20 years. Moore's law illustrates this and indicates a continuing evolution of processor performance and increased memory size, often combined with reduced size, power consumption, and cost for devices. High-resolution color displays and megapixel camera sensors are also coming into all types of mobile devices. Combined with a high-speed internet backbone often based on optical fiber networks, we see that a range of technology enablers are in place to go hand-in-hand with advancement in mobile communications technology such as LTE.

The rapid increase in use of the internet to provide all kinds of services since the 1990s started at the same time as 2G and 3G mobile systems came into widespread use. The natural next step was that those internet-based services also moved to the mobile devices, creating what is today known as *mobile broadband*. Being able to support the same *Internet Protocol* (IP)-based services in a mobile device that people use at home with a fixed broadband connection is a major challenge and a prime driver for the evolution of LTE. A few services were already supported by the evolved 2.5G systems, but it is not until the systems are designed primarily for IP-based services that the real mobile IP revolution can take off. An interesting aspect of the migration of broadband services to mobile devices is that a mobile "flavor" is also added. The mobile position and the mobility and roaming capabilities do in fact create a whole new range of services tailored to the mobile environment.

Fixed telephony (POTS) and earlier generations of mobile technology were built for circuit switched services, primarily voice. The first data services over GSM were circuit switched, with packet-based GPRS coming in as a later addition. This also influenced the first development of 3G, which was based on circuit switched data, with packet-switched services as an add-on. It was not until the 3G evolution into HSPA and later LTE/LTE-Advanced that packet-switched services and IP were made the primary design target. The old circuit-switched services remain, but will on LTE be provided over IP, with Voice-over IP (VoIP) as an example.

IP is in itself service agnostic and thereby enables a range of services with different requirements. The main service-related design parameters for a radio interface supporting a variety of services are:

- **Data rate.** Many services with lower data rates such as voice services are important and still occupy a large part of a mobile network's overall capacity, but it is the higher data rate services that drive the design of the radio interface. The ever increasing demand for higher data rates for web browsing, streaming and file transfer pushes the peak data rates for mobile systems from kbit/s for 2G, to Mbit/s for 3G and getting close to Gbit/s for 4G.
- **Delay.** Interactive services such as real-time gaming, but also web browsing and interactive file transfer, have requirements for very low delay, making it a primary design target. There are, however, many applications such as e-mail and television where the delay requirements are not as strict. The delay for a packet sent from a server to a client and back is called *latency*.

- **Capacity.** From the mobile system operator's point of view, it is not only the peak data rates provided to the end-user that are of importance, but also the total data rate that can be provided on average from each deployed base station site and per hertz of licensed spectrum. This measure of capacity is called *spectral efficiency*. In the case of capacity shortage in a mobile system, the Quality-of-Service (QoS) for the individual end-users may be degraded.

How these three main design parameters influenced the development of LTE is described in more detail in Chapter 7 and an evaluation of what performance is achieved for the design parameters above is presented in Chapter 18.

The demand for new services and for higher peak bit rates and system capacity is not only met by evolution of the technology to 4G. There is also a demand for more spectrum resources to expand systems and the demand also leads to more competition between an increasing number of mobile operators and between alternative technologies to provide mobile broadband services. An overview of some technologies other than LTE is given in Chapter 19.

With more spectrum coming into use for mobile broadband, there is a need to operate mobile systems in a number of different frequency bands, in spectrum allocations of different sizes and sometimes also in fragmented spectrum. This calls for high spectrum flexibility with the possibility for a varying channel bandwidth, which was also a driver and an essential design parameter for LTE.

The demand for new mobile services and the evolution of the radio interface to LTE have served as drivers to evolve the core network. The core network developed for GSM in the 1980s was extended to support GPRS, EDGE, and WCDMA in the 1990s, but was still very much built around the circuit-switched domain. A *System Architecture Evolution* (SAE) was initiated at the same time as LTE development started and has resulted in an *Evolved Packet Core* (EPC), developed to support HSPA and LTE/LTE-Advanced, focusing on the packet-switched domain. For more details on SAE/EPC, please refer to [9].

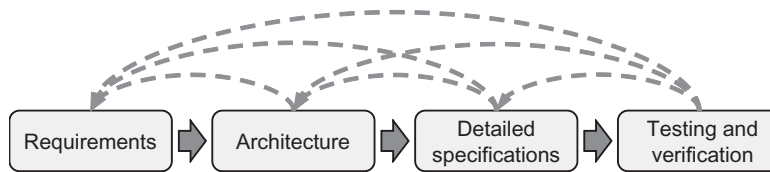
1.5 STANDARDIZATION OF LTE

With a framework for IMT systems set up by the ITU-R, with spectrum made available by the WRC and with an ever increasing demand for better performance, the task of specifying the LTE system that meets the design targets falls on 3GPP. 3GPP writes specifications for 2G, 3G, and 4G mobile systems, and 3GPP technologies are the most widely deployed in the world, with more than 4.5 billion connections in 2010. In order to understand how 3GPP works, it is important to understand the process of writing standards.

1.5.1 The Standardization Process

Setting a standard for mobile communication is not a one-time job, it is an ongoing process. The standardization forums are constantly evolving their standards trying to meet new demands for services and features. The standardization process is different in the different forums, but typically includes the four phases illustrated in [Figure 1.4](#):

1. *Requirements*, where it is decided what is to be achieved by the standard.
2. *Architecture*, where the main building blocks and interfaces are decided.

**FIGURE 1.4**

The standardization phases and iterative process.

3. *Detailed specifications*, where every interface is specified in detail.
4. *Testing and verification*, where the interface specifications are proven to work with real-life equipment.

These phases are overlapping and iterative. As an example, requirements can be added, changed, or dropped during the later phases if the technical solutions call for it. Likewise, the technical solution in the detailed specifications can change due to problems found in the testing and verification phase.

Standardization starts with the *requirements* phase, where the standards body decides what should be achieved with the standard. This phase is usually relatively short.

In the *architecture* phase, the standards body decides about the architecture – that is, the principles of how to meet the requirements. The architecture phase includes decisions about reference points and interfaces to be standardized. This phase is usually quite long and may change the requirements.

After the architecture phase, the *detailed specification* phase starts. It is in this phase the details for each of the identified interfaces are specified. During the detailed specification of the interfaces, the standards body may find that previous decisions in the architecture or even in the requirements phases need to be revisited.

Finally, the *testing and verification* phase starts. It is usually not a part of the actual standardization in the standards bodies, but takes place in parallel through testing by vendors and interoperability testing between vendors. This phase is the final proof of the standard. During the testing and verification phase, errors in the standard may still be found and those errors may change decisions in the detailed standard. Albeit not common, changes may also need to be made to the architecture or the requirements. To verify the standard, products are needed. Hence, the implementation of the products starts after (or during) the detailed specification phase. The testing and verification phase ends when there are stable test specifications that can be used to verify that the equipment is fulfilling the standard.

Normally, it takes one to two years from the time when the standard is completed until commercial products are out on the market. However, if the standard is built from scratch, it may take longer since there are no stable components to build from.

1.5.2 The 3GPP Process

The *Third-Generation Partnership Project* (3GPP) is the standards-developing body that specifies the LTE/LTE-Advanced, as well as 3G UTRA and 2G GSM systems. 3GPP is a partnership project formed by the standards bodies ETSI, ARIB, TTC, TTA, CCSA, and ATIS. 3GPP consists of four *Technical Specifications Groups* (TSGs) – see Figure 1.5.

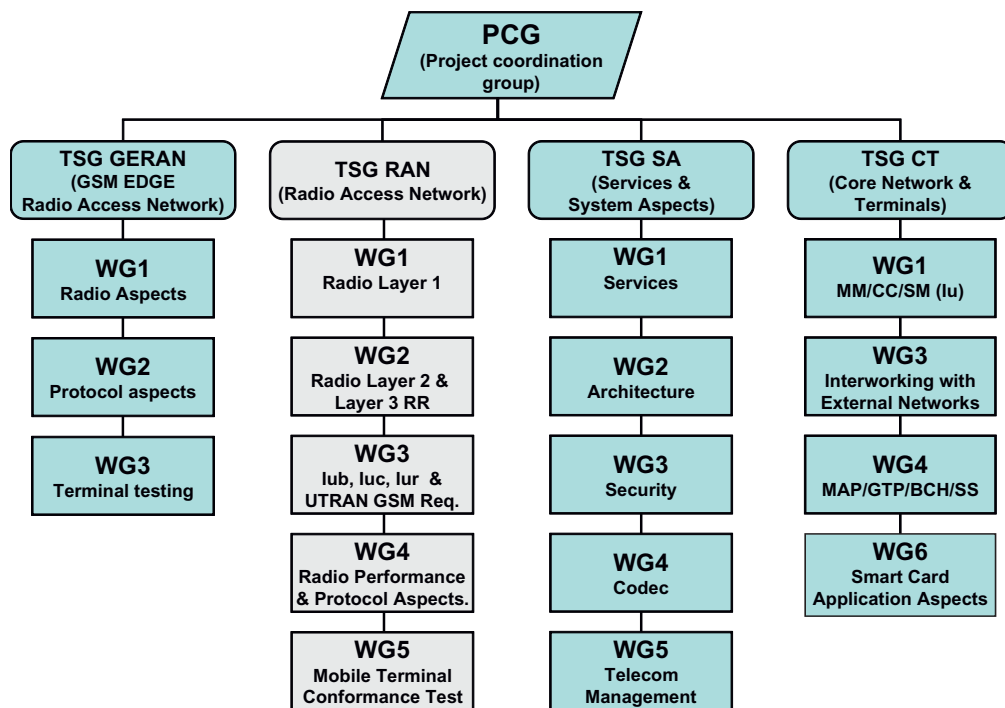


FIGURE 1.5

3GPP organization.

A parallel partnership project called 3GPP2 was formed in 1999. It also develops 3G specifications, but for CDMA2000, which is the 3G technology developed from the 2G CDMA-based standard IS-95. It is also a global project, and the organizational partners are ARIB, CCSA, TTA, and TTC.

3GPP TSG RAN (*Radio Access Network*) is the technical specification group that has developed WCDMA, its evolution HSPA, as well as LTE/LTE-Advanced, and is in the forefront of the technology. TSG RAN consists of five working groups (WGs):

1. RAN WG1, dealing with the physical layer specifications.
2. RAN WG2, dealing with the layer 2 and layer 3 radio interface specifications.
3. RAN WG3, dealing with the fixed RAN interfaces – for example, interfaces between nodes in the RAN – but also the interface between the RAN and the core network.
4. RAN WG4, dealing with the *radio frequency* (RF) and *radio resource management* (RRM) performance requirements.
5. RAN WG 5, dealing with the terminal conformance testing.

The scope of 3GPP when it was formed in 1998 was to produce global specifications for a 3G mobile system based on an evolved GSM core network, including the WCDMA-based radio access of

the UTRA FDD and the TD-CDMA-based radio access of the UTRA TDD mode. The task to maintain and develop the GSM/EDGE specifications was added to 3GPP at a later stage and the work now also includes LTE (E-UTRA). The UTRA, E-UTRA and GSM/EDGE specifications are developed, maintained, and approved in 3GPP. After approval, the organizational partners transpose them into appropriate deliverables as standards in each region.

In parallel with the initial 3GPP work, a 3G system based on TD-SCDMA was developed in China. TD-SCDMA was eventually merged into release 4 of the 3GPP specifications as an additional TDD mode.

The work in 3GPP is carried out with relevant ITU recommendations in mind and the result of the work is also submitted to ITU. The organizational partners are obliged to identify regional requirements that may lead to options in the standard. Examples are regional frequency bands and special protection requirements local to a region. The specifications are developed with global roaming and circulation of terminals in mind. This implies that many regional requirements in essence will be global requirements for all terminals, since a roaming terminal has to meet the strictest of all regional requirements. Regional options in the specifications are thus more common for base stations than for terminals.

The specifications of all releases can be updated after each set of TSG meetings, which occur four times a year. The 3GPP documents are divided into releases, where each release has a set of features added compared to the previous release. The features are defined in Work Items agreed and undertaken by the TSGs. The releases from release 8 and onwards, with some main features listed for LTE, are shown in Figure 1.6. The date shown for each release is the day the content of the release was frozen. Release 10 of LTE is the version approved by ITU-R as an IMT-Advanced technology and is therefore also named *LTE-Advanced*.

1.5.3 The 3G Evolution to 4G

The first release of WCDMA Radio Access developed in TSG RAN was called release 99³ and contained all features needed to meet the IMT-2000 requirements as defined by the ITU. This included circuit-switched voice and video services, and data services over both packet-switched and circuit-switched bearers. The first major addition of radio access features to WCDMA was HSPA, which was added in release 5 with *High Speed Downlink Packet Access* (HSDPA) and release 6 with *Enhanced Uplink*. These two are together referred to as HSPA and an overview of HSPA is given in Chapter 19 of this book. With HSPA, UTRA goes beyond the definition of a 3G mobile system and also encompasses broadband mobile data.

The 3G evolution continued in 2004, when a workshop was organized to initiate work on the 3GPP *Long-Term Evolution* (LTE) radio interface. The result of the LTE workshop was that a study item in 3GPP TSG RAN was created in December 2004. The first 6 months were spent on defining the requirements, or design targets, for LTE. These were documented in a 3GPP technical report [10] and approved in June 2005. Most notable are the requirements on high data rate at the cell edge and the importance of low delay, in addition to the normal capacity and peak data rate requirements. Furthermore, spectrum flexibility and maximum commonality between FDD and TDD solutions are pronounced.

³For historical reasons, the first 3GPP release is named after the year it was frozen (1999), while the following releases are numbered 4, 5, 6, etc.

During the fall of 2005, 3GPP TSG RAN WG1 made extensive studies of different basic physical layer technologies and in December 2005 the TSG RAN plenary decided that the LTE radio access should be based on OFDM in the downlink and DFT-precoded OFDM in the uplink. TSG RAN and its working groups then worked on the LTE specifications and the specifications were approved in December 2007. Work has since then continued on LTE, with new features added in each release, as shown in Figure 1.6. Chapters 7–17 will go through the details of the LTE radio interface in more detail.

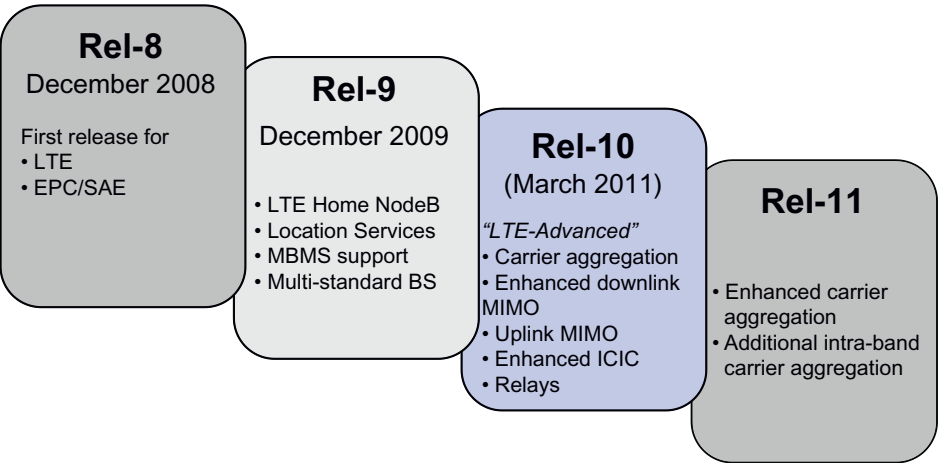


FIGURE 1.6
Releases of 3GPP specifications for LTE.

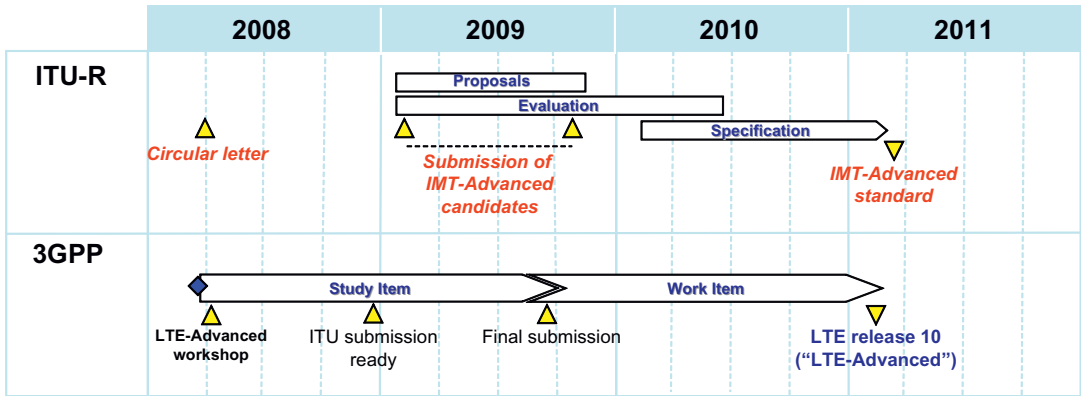


FIGURE 1.7
3GPP time schedule for LTE-Advanced in relation to ITU time-schedule on IMT-Advanced.

The work on IMT-Advanced within ITU-R WP5D came in 2008 into a phase where the detailed requirements and process were announced through a circular letter [5]. Among other things, this triggered activities in 3GPP, where a study item on *LTE-Advanced* was started. The task was to define requirements and investigate and propose technology components to be part of LTE-Advanced. The work was turned into a Work Item in 2009 in order to develop the detailed specifications.

Within 3GPP, LTE-Advanced is seen as the next major step in the evolution of LTE. LTE-Advanced is therefore not a new technology; it is an evolutionary step in the continuing development of LTE. As shown in Figure 1.6, the features that form LTE-Advanced are part of release 10 of 3GPP LTE specifications. Wider bandwidth through aggregation of multiple carriers and evolved use of advanced antenna techniques in both uplink and downlink are the major components added in LTE release 10 to reach the IMT-Advanced targets.

The work on LTE-Advanced within 3GPP is planned with the ITU-R time frame in mind, as shown in Figure 1.7. LTE-Advanced was submitted as a candidate to the ITU-R in 2009 and is now included in the set of radio interface technologies announced by ITU-R [6] in October 2010 to be included as a part of the IMT-Advanced radio interface specifications. This is very much aligned with what was from the start stated as a goal for LTE, namely that *LTE should provide the starting point for a smooth transition to 4G (= IMT-Advanced) radio access*.

Since LTE-Advanced is an integral part of 3GPP LTE release 10, it is described in detail together with the corresponding components of LTE in Chapters 7–17 of this book.

High Data Rates in Mobile Communication

As discussed in Chapter 1, one main target for the evolution of mobile communication is to provide the possibility for significantly higher end-user data rates compared to what is achievable with, for example, the first releases of the 3G standards. This includes the possibility for higher peak data rates but, as pointed out in the previous chapter, even more so the possibility for significantly higher data rates over the entire cell area, also including, for example, users at the cell edge. The initial part of this chapter will briefly discuss some of the more fundamental constraints that exist in terms of what data rates can actually be achieved in different scenarios. This will provide a background to subsequent discussions in the later part of the chapter, as well as in the subsequent chapters, concerning different means to increase the achievable data rates in different mobile-communication scenarios.

2.1 HIGH DATA RATES: FUNDAMENTAL CONSTRAINTS

In Ref. [11], Shannon provided the basic theoretical tools needed to determine the maximum rate, also known as the *channel capacity*, by which information can be transferred over a given communication channel. Although relatively complicated in the general case, for the special case of communication over a channel, for example a radio link, only impaired by additive white Gaussian noise, the channel capacity C is given by the relatively simple expression [12]:

$$C = BW \cdot \log_2 \left(1 + \frac{S}{N} \right), \quad (2.1)$$

where BW is the bandwidth available for the communication, S denotes the received signal power, and N denotes the power of the white noise impairing the received signal.

Already from Eqn (2.1) it should be clear that the two fundamental factors limiting the achievable data rate are the available received signal power, or more generally the available signal-power-to-noise-power ratio S/N , and the available bandwidth. To further clarify how and when these factors limit the achievable data rate, assume communication with a certain information rate R . The received signal power can then be expressed as $S = E_b \cdot R$, where E_b is the received energy per information bit. Furthermore, the noise power can be expressed as $N = N_0 \cdot BW$, where N_0 is the constant noise power spectral density measured in W/Hz.

Clearly, the information rate can never exceed the channel capacity. Together with the above expressions for the received signal power and noise power, this leads to the inequality:

$$R \leq C = BW \cdot \log_2 \left(1 + \frac{S}{N} \right) = BW \cdot \log_2 \left(1 + \frac{E_b \cdot R}{N_0 \cdot BW} \right) \quad (2.2)$$

or, by defining the radio-link *bandwidth utilization* $\gamma = R/BW$,

$$\gamma \leq \log_2 \left(1 + \gamma \cdot \frac{E_b}{N_0} \right). \quad (2.3)$$

This inequality can be reformulated to provide a lower bound on the required received energy per information bit, normalized to the noise power density, for a given bandwidth utilization γ :

$$\frac{E_b}{N_0} \geq \min \left\{ \frac{E_b}{N_0} \right\} = \frac{2^\gamma - 1}{\gamma}. \quad (2.4)$$

The rightmost expression – that is, the minimum required E_b/N_0 at the receiver as a function of the bandwidth utilization – is illustrated in Figure 2.1. As can be seen, for bandwidth utilizations significantly less than 1 – that is, for information rates substantially smaller than the utilized bandwidth – the minimum required E_b/N_0 is relatively constant, regardless of γ . For a given noise power density, any increase of the information data rate then implies a similar relative increase in the minimum required signal power $S = E_b \cdot R$ at the receiver. On the other hand, for bandwidth utilizations larger than 1, the minimum required E_b/N_0 increases rapidly with γ . Thus, in the case of data rates of the same order as or larger than the communication bandwidth, any further increase of the information data rate, without a corresponding increase in the available bandwidth, implies a larger, eventually much larger, relative increase in the minimum required received signal power.

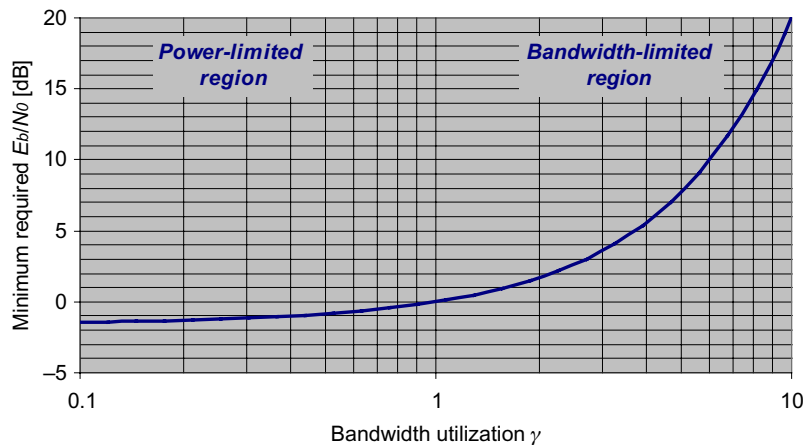


FIGURE 2.1

Minimum required E_b/N_0 at the receiver as a function of bandwidth utilization.

2.1.1 High Data Rates in Noise-Limited Scenarios

From the discussion above, some basic conclusions can be drawn regarding the provisioning of higher data rates in a mobile-communication system when noise is the main source of radio-link impairment (a *noise-limited* scenario):

- The data rates that can be provided in such scenarios are always limited by the available received signal power or, in the general case, the received signal-power-to-noise-power ratio. Furthermore, any increase of the achievable data rate within a given bandwidth will require *at least* the same relative increase of the received signal power. At the same time, if sufficient received signal power can be made available, basically any data rate can, at least in theory, be provided within a given limited bandwidth.
- In the case of low-bandwidth utilization – that is, as long as the radio-link data rate is substantially lower than the available bandwidth – any further increase of the data rate requires *approximately the same* relative increase in the received signal power. This can be referred to as *power-limited* operation (in contrast to *bandwidth-limited* operation; see below) as, in this case, an increase in the available bandwidth does not substantially impact what received signal power is required for a certain data rate.
- On the other hand, in the case of high-bandwidth utilization – that is, in the case of data rates of the same order as or exceeding the available bandwidth – any further increase in the data rate requires a *much larger* relative increase in the received signal power unless the bandwidth is increased in proportion to the increase in data rate. This can be referred to as *bandwidth-limited operation* as, in this case, an increase in the bandwidth will reduce the received signal power required for a certain data rate.

Thus, to make efficient use of the available received signal power or, in the general case, the available signal-to-noise ratio, the transmission bandwidth should at least be of the same order as the data rates to be provided.

Assuming a constant transmit power, the received signal power can always be increased by reducing the distance between the transmitter and the receiver, thereby reducing the attenuation of the signal as it propagates from the transmitter to the receiver. Thus, in a noise-limited scenario it is, at least in theory, always possible to increase the achievable data rates, assuming that one is prepared to accept a reduction in the transmitter/receiver distance – that is, a reduced range. In a mobile-communication system this would correspond to a reduced cell size and thus the need for more cell sites to cover the same overall area. In particular, providing data rates of the same order as or larger than the available bandwidth – that is, with a high-bandwidth utilization – would require a significant cell-size reduction. Alternatively, one has to accept that the high data rates are only available for terminals in the center of the cell and not over the entire cell area.

Another means to increase the overall received signal power for a given transmit power is the use of additional antennas at the receiver side, also known as *receive-antenna diversity*. Multiple receive antennas can be applied at the base station (that is, for the uplink) or at the terminal (that is, for the downlink). By proper combination of the signals received at the different antennas, the signal-to-noise ratio after the antenna combination can be increased in proportion to the number of receive antennas, thereby allowing for higher data rates for a given transmitter/receiver distance.

Multiple antennas can also be applied at the transmitter side, typically at the base station, and be used to focus a given total transmit power in the direction of the receiver – that is, toward the target terminal. This will increase the received signal power and thus, once again, allow for higher data rates for a given transmitter/receiver distance.

However, providing higher data rates by the use of multiple transmit or receive antennas is only efficient up to a certain level – that is, as long as the data rates are power limited rather than bandwidth limited. Beyond this point, the achievable data rates start to saturate and any further increase in the number of transmit or receive antennas, although leading to a correspondingly improved signal-to-noise ratio at the receiver, will only provide a marginal increase in the achievable data rates. This saturation in achievable data rates can be avoided though, by the use of multiple antennas at both the transmitter *and* the receiver, enabling what can be referred to as *spatial multiplexing*, often also referred to as *MIMO* (Multiple-Input Multiple-Output). Different types of multi-antenna techniques, including spatial multiplexing, will be discussed in more detail in Chapter 5. Multi-antenna techniques for the specific case of LTE are discussed in Chapters 10 and 11.

An alternative to increasing the received signal power is to reduce the noise power, or more exactly the noise power density, at the receiver. This can, at least to some extent, be achieved by more advanced receiver RF design, allowing for a reduced receiver noise figure.

2.1.2 Higher Data Rates in Interference-Limited Scenarios

The discussion above assumed communication over a radio link only impaired by noise. However, in actual mobile-communication scenarios, interference from transmissions in neighboring cells, also referred to as *inter-cell interference*, is often the dominant source of radio-link impairment, more so than noise. This is especially the case in small-cell deployments with a high traffic load. Furthermore, in addition to inter-cell interference there may in some cases also be interference from other transmissions *within the current cell*, also referred to as *intra-cell interference*.

In many respects the impact of interference on a radio link is similar to that of noise. In particular, the basic principles discussed above apply also to a scenario where interference is the main radio-link impairment:

- The maximum data rate that can be achieved in a given bandwidth is limited by the available signal-power-to-interference-power ratio.
- Providing data rates larger than the available bandwidth (high-bandwidth utilization) is costly in the sense that it requires a disproportionately high signal-to-interference ratio.

Also, similar to a scenario where noise is the dominant radio-link impairment, reducing the cell size as well as the use of multi-antenna techniques are key means to increase the achievable data rates in an interference-limited scenario:

- Reducing the cell size will obviously reduce the number of users, and thus also the overall traffic, per cell. This will reduce the relative interference level and thus allow for higher data rates.
- Similar to the increase in signal-to-noise ratio, proper combination of the signals received at multiple antennas will also increase the signal-to-interference ratio after the antenna combination.
- The use of beam-forming by means of multiple transmit antennas will focus the transmit power in the direction of the target receiver, leading to reduced interference to other radio links and thus improving the overall signal-to-interference ratio in the system.

One important difference between interference and noise is that interference, in contrast to noise, typically has a certain structure which makes it, at least to some extent, predictable and thus possible to further suppress or even remove completely. As an example, a dominant interfering signal

may arrive from a certain direction, in which case the corresponding interference can be further suppressed, or even completely removed, by means of *spatial processing* using multiple antennas at the receiver. This will be further discussed in Chapter 5. Also, any differences in the spectral properties between the target signal and an interfering signal can be used to suppress the interferer and thus reduce the overall interference level.

2.2 HIGHER DATA RATES WITHIN A LIMITED BANDWIDTH: HIGHER-ORDER MODULATION

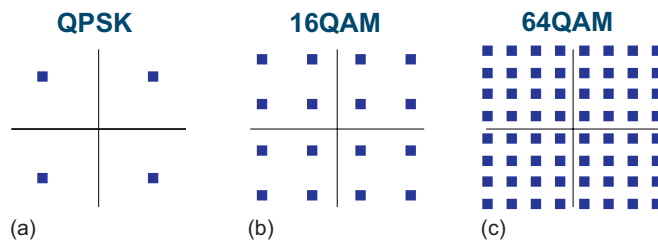
As discussed in the previous section, providing data rates larger than the available bandwidth is fundamentally inefficient in the sense that it requires disproportionately high signal-to-noise and signal-to-interference ratios at the receiver. Still, bandwidth is often a scarce and expensive resource and, at least in some mobile-communication scenarios, high signal-to-noise and signal-to-interference ratios can be made available, for example in small-cell environments with a low traffic load or for terminals close to the cell site. Mobile-communication systems should preferably be designed to be able to take advantage of such scenarios – that is, they should be able to offer very high data rates within a limited bandwidth when the radio conditions so allow.

A straightforward means to provide higher data rates within a given transmission bandwidth is the use of *higher-order modulation*, implying that the modulation alphabet is extended to include additional signaling alternatives and thus allowing for more bits of information to be communicated per modulation symbol.

In the case of QPSK modulation, which is the modulation scheme used for the downlink in the first releases of the 3G mobile-communication standards (WCDMA and CDMA2000), the modulation alphabet consists of four different signaling alternatives. These four signaling alternatives can be illustrated as four different points in a two-dimensional plane (see Figure 2.2a). With four different signaling alternatives, QPSK allows for up to 2 bits of information to be communicated during each modulation-symbol interval. By extending to 16QAM modulation (Figure 2.2b), 16 different signaling alternatives are available. The use of 16QAM thus allows for up to 4 bits of information to be communicated per symbol interval. Further extension to 64QAM (Figure 2.2c), with 64 different signaling alternatives, allows for up to 6 bits of information to be communicated per symbol interval. At the same time, the bandwidth of the transmitted signal is, at least in principle, independent of the size of the modulation alphabet and mainly depends on the modulation rate – that is, the number of modulation symbols per second. The maximum bandwidth utilization, expressed in bits/s/Hz, of 16QAM and 64QAM are thus, at least in principle, two and three times that of QPSK respectively.

It should be pointed out that there are many other possible modulation schemes, in addition to those illustrated in Figure 2.2. One example is 8PSK, consisting of eight signaling alternatives and thus providing up to 3 bits of information per modulation symbol. Readers are referred to [12] for a more thorough discussion on different modulation schemes.

The use of higher-order modulation provides the possibility for higher bandwidth utilization – that is, the possibility to provide higher data rates within a given bandwidth. However, the higher bandwidth utilization comes at the cost of reduced robustness to noise and interference. Alternatively expressed, higher-order modulation schemes, such as 16QAM or 64QAM, require a higher E_b/N_0 at the receiver for a given bit-error probability, compared to QPSK. This is in line with the discussion in

**FIGURE 2.2**

Signal constellations for: (a) QPSK; (b) 16QAM; (c) 64QAM.

the previous section, where it was concluded that high-bandwidth utilization – that is, a high information rate within a limited bandwidth – in general requires a higher receiver E_b/N_0 .

2.2.1 Higher-Order Modulation in Combination with Channel Coding

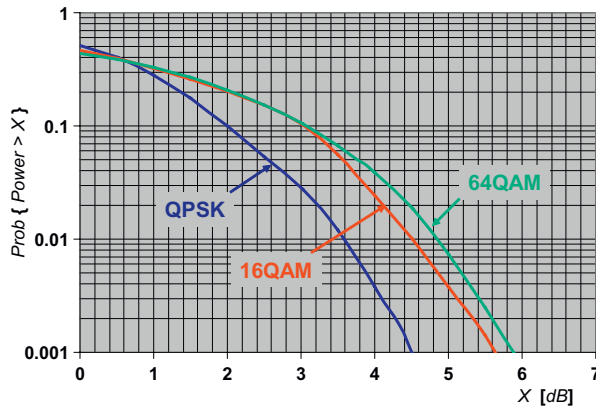
Higher-order modulation schemes such as 16QAM and 64QAM require, in themselves, a higher receiver E_b/N_0 for a given error rate, compared to QPSK. However, in combination with channel coding the use of higher-order modulation will sometimes be more efficient – that is, require a lower receiver E_b/N_0 for a given error rate – compared to the use of lower-order modulation such as QPSK. This may, for example, occur when the target bandwidth utilization implies that, with lower-order modulation, no or very little channel coding can be applied. In such a case, the additional channel coding that can be applied by using a higher-order modulation scheme such as 16QAM may lead to an overall gain in power efficiency compared to the use of QPSK.

As an example, if a bandwidth utilization of close to 2 information bits per modulation symbol is required, QPSK modulation would allow for very limited channel coding (channel-coding rate close to 1). On the other hand, the use of 16QAM modulation would allow for a channel-coding rate of the order of one-half. Similarly, if a bandwidth efficiency close to 4 information bits per modulation symbol is required, the use of 64QAM may be more efficient than 16QAM modulation, taking into account the possibility for lower-rate channel coding and corresponding additional coding gain in the case of 64QAM. It should be noted that this does not contradict the general discussion in Section 2.1, where it was concluded that transmission with high-bandwidth utilization is inherently power inefficient. The use of rate 1/2 channel coding for 16QAM obviously reduces the information data rate, and thus also the bandwidth utilization, to the same level as uncoded QPSK.

From the discussion above it can be concluded that, for a given signal-to-noise/interference ratio, a certain combination of modulation scheme and channel-coding rate is optimal in the sense that it can deliver the highest-bandwidth utilization (the highest data rate within a given bandwidth) for that signal-to-noise/interference ratio.

2.2.2 Variations in Instantaneous Transmit Power

A general drawback of higher-order modulation schemes such as 16QAM and 64QAM, where information is also encoded in the instantaneous amplitude of the modulated signal, is that the modulated signal will have larger variations, and thus also larger peaks, in its instantaneous power. This can be

**FIGURE 2.3**

Distribution of instantaneous power for different modulation schemes. Average power is the same in all cases.

seen in Figure 2.3, which illustrates the distribution of the instantaneous power, more specifically the probability that the instantaneous power is above a certain value, for QPSK, 16QAM, and 64QAM respectively. Clearly, the probability for large peaks in the instantaneous power is higher in the case of higher-order modulation.

Larger peaks in the instantaneous signal power imply that the transmitter power amplifier must be over-dimensioned to avoid power-amplifier nonlinearities, occurring at high instantaneous power levels, causing corruption to the signal to be transmitted. As a consequence, the power-amplifier efficiency will be reduced, leading to increased power consumption. In addition, there will be a negative impact on the power-amplifier cost. Alternatively, the average transmit power must be reduced, implying a reduced range for a given data rate. High power-amplifier efficiency is especially important for the terminal – that is, in the uplink direction – due to the importance of low mobile-terminal power consumption and cost. For the base station, high power-amplifier efficiency, although far from irrelevant, is still somewhat less important. Thus, large peaks in the instantaneous signal power are less of an issue for the downlink compared to the uplink and, consequently, higher-order modulation is more suitable for the downlink compared to the uplink.

2.3 WIDER BANDWIDTH INCLUDING MULTI-CARRIER TRANSMISSION

As was shown in Section 2.1, transmission with a high-bandwidth utilization is fundamentally power inefficient in the sense that it will require disproportionately high signal-to-noise and signal-to-interference ratios for a given data rate. Providing very high data rates within a limited bandwidth, for example by means of higher-order modulation, is thus only possible in situations where relatively high signal-to-noise and signal-to-interference ratios can be made available, for example in small-cell environments with low traffic load or for terminals close to the cell site.

Instead, to provide high data rates as efficiently as possible in terms of required signal-to-noise and signal-to-interference ratios, implying as good coverage as possible for high data rates, the transmission bandwidth should be at least of the same order as the data rates to be provided.

Having in mind that the provisioning of higher data rates with good coverage is one of the main targets for mobile communication, it can thus be concluded that support for even wider transmission bandwidth is an important part of this evolution.

However, there are several critical issues related to the use of wider transmission bandwidths in a mobile-communication system:

- Spectrum is, as already mentioned, often a scarce and expensive resource, and it may be difficult to find spectrum allocations of sufficient size to allow for very wideband transmission, especially at lower-frequency bands.
- The use of wider transmission and reception bandwidths has an impact on the complexity of the radio equipment, both at the base station and at the terminal. As an example, a wider transmission bandwidth has a direct impact on the transmitter and the receiver sampling rates, and thus on the complexity and power consumption of digital-to-analog and analog-to-digital converters, as well as front-end digital signal processing. RF components are also, in general, more complicated to design and more expensive to produce, the wider the bandwidth they have to handle.

The two issues above are mainly outside the scope of this book. However, a more specific technical issue related to wider-band transmission is the increased corruption of the transmitted signal due to time dispersion on the radio channel. Time dispersion occurs when the transmitted signal propagates to the receiver via multiple paths with different delays (see Figure 2.4a). In the frequency domain, a time-dispersive channel corresponds to a non-constant channel frequency response, as illustrated in Figure 2.4b. This radio-channel *frequency selectivity* will corrupt the frequency-domain structure of the transmitted signal and lead to higher error rates for given signal-to-noise/interference ratios. Every radio channel is subject to frequency selectivity, at least to some extent. However, the extent to which the frequency selectivity impacts the radio communication depends on the bandwidth of the transmitted signal with, in general, larger impact for wider-band transmission. The amount of radio-channel frequency selectivity also depends on the environment, with typically less frequency selectivity (less time dispersion) in the case of small cells and in environments with few obstructions and potential reflectors, such as rural environments.

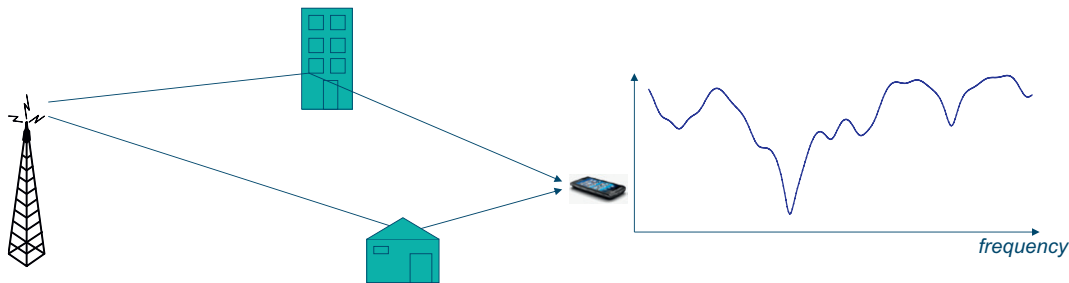


FIGURE 2.4

Multi-path propagation causing time dispersion and radio-channel frequency selectivity.

It should be noted that Figure 2.4b illustrates a “snapshot” of the channel frequency response. As a terminal is moving through the environment, the detailed structure of the multi-path propagation, and thus also the detailed structure of the channel frequency response, may vary rapidly with time. The rate of the variations in the channel frequency response is related to the channel *Doppler spread*, f_D , defined as $f_D = v/c \cdot f_c$, where v is the speed of the terminal, f_c is the carrier frequency (for example, 2 GHz), and c is the speed of light.

Receiver-side *equalization* [12] has for many years been used to counteract signal corruption due to radio-channel frequency selectivity. Equalization has been shown to provide satisfactory performance with reasonable complexity at least up to bandwidths corresponding to the WCDMA bandwidth of 5 MHz (see, for example, [13]). However, if the transmission bandwidth is further increased up to, for example, 20 MHz, as is done in LTE, the complexity of straightforward high-performance equalization starts to become a serious issue. One option is then to apply less optimal equalization, with a corresponding negative impact on the equalizer capability to counteract the signal corruption due to radio-channel frequency selectivity and thus a corresponding negative impact on the radio-link performance.

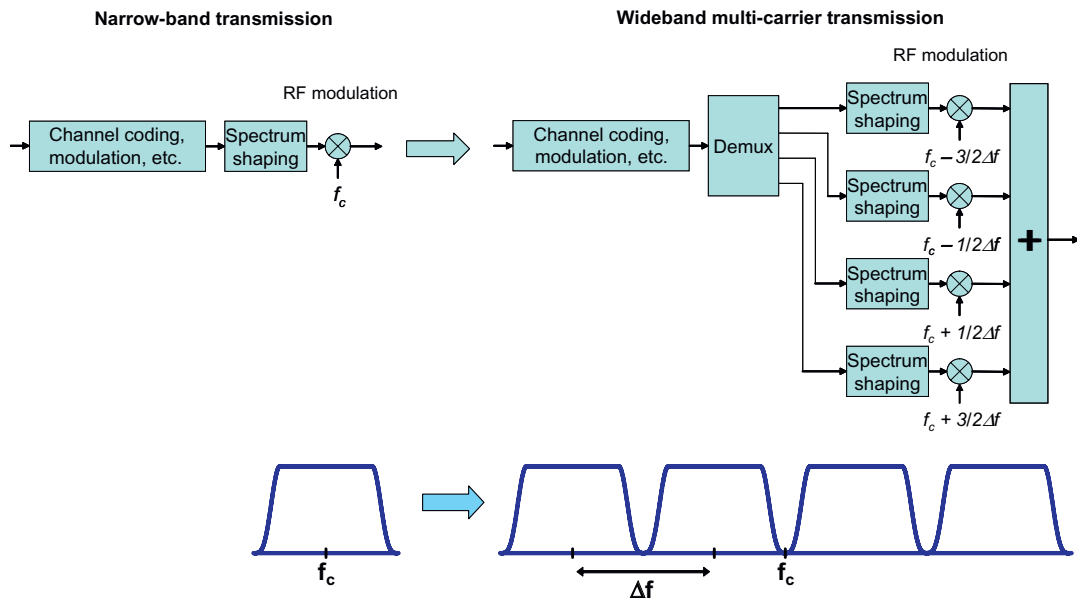
An alternative approach is to consider specific transmission schemes and signal designs that allow for good radio-link performance also in the case of substantial radio-channel frequency selectivity without a prohibitively large receiver complexity. In the following, two such approaches to wider-band transmission will be discussed:

1. The use of different types of *multi-carrier transmission* – that is, transmitting an overall wider-band signal as several more narrowband frequency-multiplexed signals (see below). One special case of multi-carrier transmission is *OFDM transmission*, to be discussed in more detail in Chapter 3.
2. The use of specific *single-carrier* transmission schemes, especially designed to allow for efficient but still reasonably low-complexity equalization. This is further discussed in Chapter 4.

2.3.1 Multi-Carrier Transmission

One way to increase the overall transmission bandwidth, without suffering from increased signal corruption due to radio-channel frequency selectivity, is the use of so-called *multi-carrier transmission*. As illustrated in Figure 2.5, multi-carrier transmission implies that, instead of transmitting a single more wideband signal, multiple more narrowband signals, often referred to as *subcarriers*, are frequency multiplexed and jointly transmitted over the same radio link to the same receiver. By transmitting M signals in parallel over the same radio link, the overall data rate can be increased up to M times. At the same time, the impact in terms of signal corruption due to radio-channel frequency selectivity depends on the bandwidth of each subcarrier. Thus, the impact from a frequency-selective channel is essentially the same as for a more narrowband transmission scheme with a bandwidth that corresponds to the bandwidth of each subcarrier.

A drawback of the kind of multi-carrier evolution outlined in Figure 2.5, where an existing more narrowband radio-access technology is extended to a wider overall transmission bandwidth by the parallel transmission of M more narrowband carriers, is that the spectrum of each subcarrier typically does not allow for very tight subcarrier “packing”. This is illustrated by the “valleys” in the overall multi-carrier spectrum outlined in the lower part of Figure 2.5. This has a somewhat negative impact on the overall bandwidth efficiency of this kind of multi-carrier transmission.

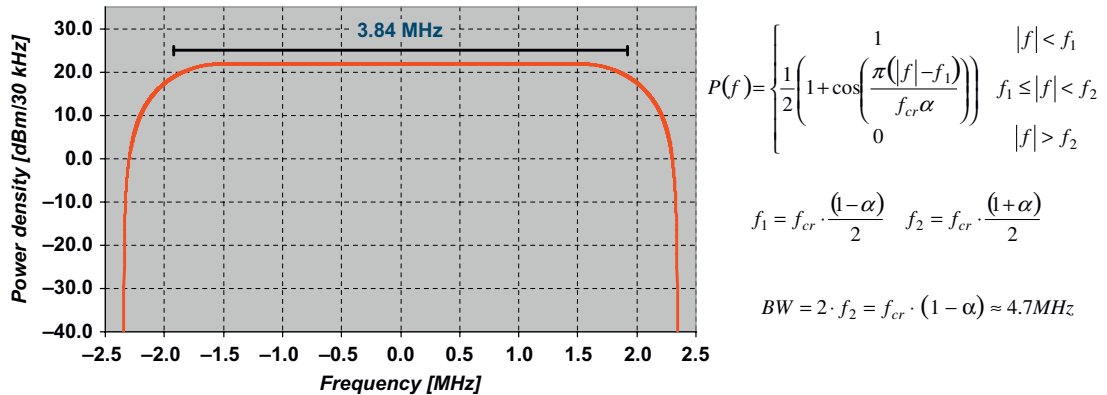
**FIGURE 2.5**

Extension to wider transmission bandwidth by means of multi-carrier transmission.

As an example, consider the multi-carrier evolution of WCDMA/HSPA towards wider bandwidth. WCDMA has a modulation rate, also referred to as the *WCDMA chip rate*, of $f_{cr} = 3.84$ Mchips/s. However, due to spectrum shaping, even the theoretical WCDMA spectrum, not including spectrum widening due to transmitter imperfections, has a bandwidth that significantly exceeds 3.84 MHz. More specifically, as can be seen in Figure 2.6, the theoretical WCDMA spectrum has a *raised-cosine* shape with roll-off $\alpha = 0.22$. As a consequence, the bandwidth outside of which the WCDMA theoretical spectrum equals zero is approximately 4.7 MHz (see right part of Figure 2.6).

For a straightforward multi-carrier extension of WCDMA, the subcarriers must thus be spaced approximately 4.7 MHz from each other to completely avoid inter-subcarrier interference. It should be noted, though, that a smaller subcarrier spacing can be used with only limited inter-subcarrier interference.

A second drawback of multi-carrier transmission is that, similar to the use of higher-order modulation, the parallel transmission of multiple carriers will lead to larger variations in the instantaneous transmit power. Thus, similar to the use of higher-order modulation, multi-carrier transmission will have a negative impact on the transmitter power-amplifier efficiency, implying increased transmitter power consumption and increased power-amplifier cost. Alternatively, the average transmit power must be reduced, implying a reduced range for a given data rate. For this reason, similar to the use of higher-order modulation, multi-carrier transmission is more suitable for the downlink (base-station transmission), compared to the uplink (mobile-terminal transmission), due to the higher importance of high power-amplifier efficiency at the terminal.

**FIGURE 2.6**

Theoretical WCDMA spectrum. Raised-cosine shape with roll-off $\alpha = 0.22$.

The main advantage with the kind of multi-carrier extension outlined in Figure 2.5 is that it provides a very smooth evolution, in terms of both radio equipment and spectrum, of an already existing radio-access technology to wider transmission bandwidth and a corresponding possibility for higher data rates, especially for the downlink. In essence this kind of multi-carrier evolution to wider bandwidth can be designed so that, for legacy terminals not capable of multi-carrier reception, each downlink “subcarrier” will appear as an original, more narrowband carrier, while, for a multi-carrier-capable terminal, the network can make use of the full multi-carrier bandwidth to provide higher data rates.

The next chapter will discuss, in more detail, a different approach to multi-carrier transmission, based on the so-called OFDM technique.

OFDM Transmission

In this chapter, a more detailed overview of OFDM or *Orthogonal Frequency-Division Multiplexing* will be given. OFDM is the transmission scheme used for 3GPP LTE and is also used for several other radio-access technologies, for example WiMAX [14] and the DVB broadcast technologies [15].

3.1 BASIC PRINCIPLES OF OFDM

Transmission by means of OFDM can be seen as a kind of multi-carrier transmission. The basic characteristics of OFDM transmission, which distinguish it from a straightforward multi-carrier extension of a more narrowband transmission scheme as outlined in Figure 2.5 in the previous chapter, are:

- The use of a typically very large number of relatively narrowband subcarriers. In contrast, a straightforward multi-carrier extension as outlined in Figure 2.5 would typically consist of only a few subcarriers, each with a relatively wide bandwidth. As an example, the HSPA multi-carrier evolution to a 20MHz overall transmission bandwidth consists of four (sub)carriers, each with a bandwidth of the order of 5 MHz. In comparison, OFDM transmission may imply that several hundred subcarriers are transmitted over the same radio link to the same receiver.
- Simple rectangular pulse shaping is illustrated in Figure 3.1a. This corresponds to a sinc-square-shaped per-subcarrier spectrum, as illustrated in Figure 3.1b.
- Tight frequency-domain packing of the subcarriers with a subcarrier spacing $\Delta f = 1/T_u$, where T_u is the per-subcarrier modulation-symbol time (see Figure 3.2). The subcarrier spacing is thus equal to the per-subcarrier modulation rate $1/T_u$.¹

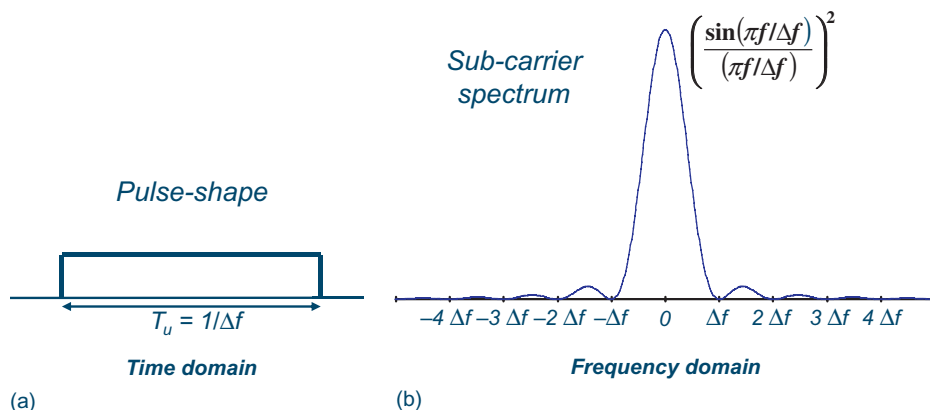
An illustrative description of a basic OFDM modulator is provided in Figure 3.3. It consists of a bank of N_c complex modulators, where each modulator corresponds to one OFDM subcarrier.

In complex baseband notation, a basic OFDM signal $x(t)$ during the time interval $mT_u \leq t < (m+1)T_u$ can thus be expressed as:

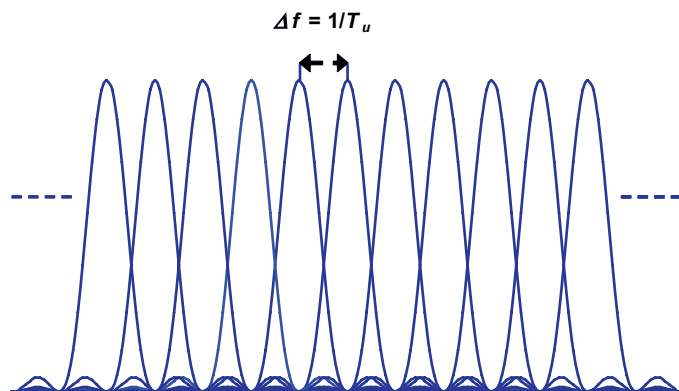
$$x(t) = \sum_{k=0}^{N_c-1} x_k(t) = \sum_{k=0}^{N_c-1} a_k^{(m)} e^{j2\pi k \Delta f t}, \quad (3.1)$$

where $x_k(t)$ is the k th modulated subcarrier with frequency $f_k = k \cdot \Delta f$ and $a_k^{(m)}$ is the, generally complex, modulation symbol applied to the k th subcarrier during the m th OFDM symbol interval – that is, during the time interval $mT_u \leq t < (m+1)T_u$. OFDM transmission is thus block based, implying

¹This ignores a possible *cyclic prefix*, see Section 3.4.

**FIGURE 3.1**

(a) Per-subcarrier pulse shape. (b) Spectrum for basic OFDM transmission.

**FIGURE 3.2**

OFDM subcarrier spacing.

that, during each OFDM symbol interval, N_c modulation symbols are transmitted in parallel. The modulation symbols can be from any modulation alphabet, such as QPSK, 16QAM, or 64QAM.

The number of OFDM subcarriers can range from less than hundred to several thousand, with the subcarrier spacing ranging from several hundred kHz down to a few kHz. What subcarrier spacing to use depends on what types of environments the system is to operate in, including such aspects as the maximum expected radio-channel frequency selectivity (maximum expected time dispersion) and the maximum expected rate of channel variations (maximum expected Doppler spread). Once the subcarrier spacing has been selected, the number of subcarriers can be decided based on the assumed overall transmission bandwidth, taking into account acceptable out-of-band emission, etc.

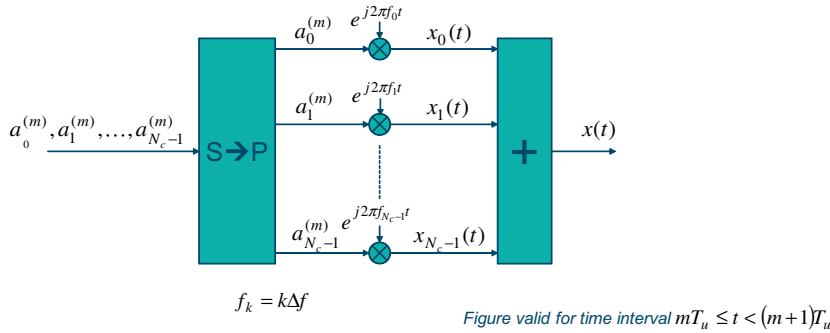


FIGURE 3.3

Principles for OFDM modulation.

The selection of OFDM subcarrier spacing and number of subcarriers is discussed in somewhat more detail in Section 3.8.

As an example, for 3GPP LTE the basic subcarrier spacing equals 15 kHz. On the other hand, the number of subcarriers depends on the transmission bandwidth, with of the order of 600 subcarriers in the case of operation in a 10 MHz spectrum allocation and correspondingly fewer/more subcarriers in the case of smaller/larger overall transmission bandwidths.

The term *Orthogonal Frequency-Division Multiplex* is due to the fact that two modulated OFDM subcarriers x_{k_1} and x_{k_2} are mutually *orthogonal* over the time interval $mT_u \leq t < (m+1)T_u$, that is:

$$\int_{mT_u}^{(m+1)T_u} x_{k_1}(t)x_{k_2}^*(t)dt = \int_{mT_u}^{(m+1)T_u} a_{k_1}a_{k_2}^*e^{j2\pi k_1\Delta ft}e^{-j2\pi k_2\Delta ft}dt = 0 \quad \text{for } k_1 \neq k_2. \quad (3.2)$$

Thus, basic OFDM transmission can be seen as the modulation of a set of orthogonal functions $\varphi_k(t)$, where

$$\varphi_k(t) = \begin{cases} e^{j2\pi k\Delta ft} & 0 \leq t < T_u \\ 0 & \text{otherwise} \end{cases}. \quad (3.3)$$

The “physical resource” in the case of OFDM transmission is often illustrated as a time–frequency grid according to Figure 3.4, where each “column” corresponds to one OFDM symbol and each “row” corresponds to one OFDM subcarrier.

3.2 OFDM DEMODULATION

Figure 3.5 illustrates the basic principle of OFDM demodulation consisting of a bank of correlators, one for each subcarrier. Taking into account the orthogonality between subcarriers according to Eqn (3.2), it is clear that, in the ideal case, two OFDM subcarriers do not cause any interference to

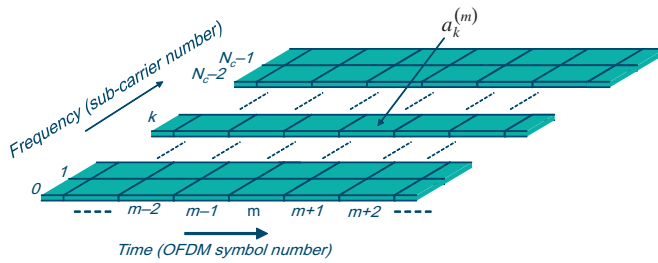


FIGURE 3.4

OFDM time–frequency grid.

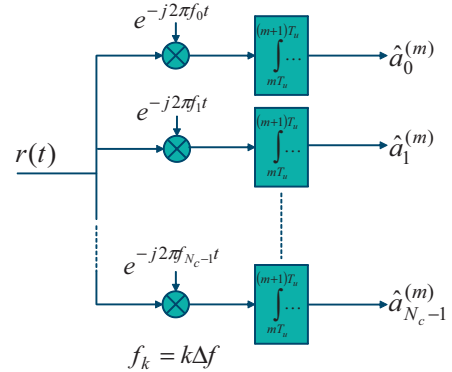


FIGURE 3.5

Basic principle of OFDM demodulation.

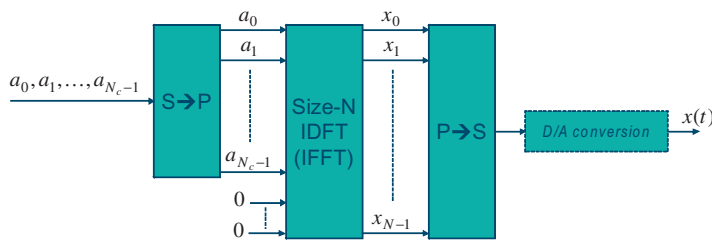
each other after demodulation. Note that this is the case despite the fact that the spectrum of neighbor subcarriers clearly overlaps, as can be seen from Figure 3.2. Thus, the avoidance of interference between OFDM subcarriers is not simply due to a subcarrier spectrum separation, which is, for example, the case for the kind of straightforward multi-carrier extension outlined in Figure 2.5 in the previous chapter. Rather, the subcarrier orthogonality is due to the *specific* frequency-domain structure of each subcarrier in combination with the *specific* choice of a subcarrier spacing Δf equal to the per-subcarrier symbol rate $1/T_u$. However, this also implies that, in contrast to the kind of multi-carrier transmission outlined in Section 2.3.1 of the previous chapter, any corruption of the frequency-domain structure of the OFDM subcarriers, for example due to a frequency-selective radio channel, may lead to a loss of inter-subcarrier orthogonality and thus to interference between subcarriers. To handle this and to make an OFDM signal truly robust to radio-channel frequency selectivity, *cyclic-prefix insertion* is typically used, as will be further discussed in Section 3.4.

3.3 OFDM IMPLEMENTATION USING IFFT/FFT PROCESSING

Although a bank of modulators/correlators according to Figures 3.3 and 3.5 can be used to illustrate the basic principles of OFDM modulation and demodulation respectively, these are not the most appropriate modulator/demodulator structures for actual implementation. Actually, due to its specific structure and the selection of a subcarrier spacing Δf equal to the per-subcarrier symbol rate $1/T_u$, OFDM allows for low-complexity implementation by means of computationally efficient *Fast Fourier Transform* (FFT) processing.

To confirm this, consider a time-discrete (sampled) OFDM signal where it is assumed that the sampling rate f_s is a multiple of the subcarrier spacing Δf – that is, $f_s = 1/T_s = N \cdot \Delta f$. The parameter N should be chosen so that the sampling theorem [12] is sufficiently fulfilled.² As $N_c \cdot \Delta f$ can be seen

²An OFDM signal defined according to Eqn (3.1) in theory has an infinite bandwidth and thus the sampling theorem can never be fulfilled completely.

**FIGURE 3.6**

OFDM modulation by means of IFFT processing.

as the nominal bandwidth of the OFDM signal, this implies that N should exceed N_c with a sufficient margin.

With these assumptions, the time-discrete OFDM signal can be expressed as:³

$$x_n = x(nT_s) = \sum_{k=0}^{N_c-1} a_k e^{j2\pi k \Delta f n T_s} = \sum_{k=0}^{N_c-1} a_k e^{j2\pi k n / N} = \sum_{k=0}^{N-1} a'_k e^{j2\pi k n / N}, \quad (3.4)$$

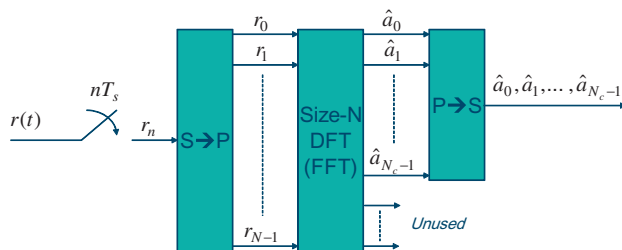
where

$$a'_k = \begin{cases} a_k & 0 \leq k < N_c \\ 0 & N_c \leq k < N \end{cases}. \quad (3.5)$$

Thus, the sequence x_n , or, in other words, the sampled OFDM signal, is the size- N Inverse Discrete Fourier Transform (IDFT) of the block of modulation symbols $a_0, a_1, \dots, a_{N_c-1}$ extended with zeros to length N . OFDM modulation can thus be implemented by means of IDFT processing followed by digital-to-analog conversion, as illustrated in Figure 3.6. In particular, by selecting the IDFT size N equal to 2^m for some integer m , the OFDM modulation can be implemented by means of implementation-efficient radix-2 Inverse Fast Fourier Transform (IFFT) processing. It should be noted that the ratio N/N_c , which could be seen as the *over-sampling* of the time-discrete OFDM signal, could very well be, and typically is, a non-integer number. As an example and as already mentioned, for 3GPP LTE the number of subcarriers N_c is approximately 600 in the case of a 10MHz spectrum allocation. The IFFT size can then, for example, be selected as $N = 1024$. This corresponds to a sampling rate $f_s = N \cdot \Delta f = 15.36\text{MHz}$, where $\Delta f = 15\text{kHz}$ is the LTE subcarrier spacing.

It is important to understand that IDFT/IFFT-based implementation of an OFDM modulator, and even more so the exact IDFT/IFFT size, are just transmitter-implementation choices and not something that would be mandated by any radio-access specification. As an example, nothing forbids the implementation of an OFDM modulator as a set of parallel modulators, as illustrated in Figure 3.3. Also, nothing prevents the use of a larger IFFT size, for example a size-2048 IFFT size, even in the case of a smaller number of OFDM subcarriers.

³From now on the index m on the modulation symbols, indicating the OFDM symbol number, will be ignored unless especially needed.

**FIGURE 3.7**

OFDM demodulation by means of FFT processing.

Similar to OFDM modulation, efficient FFT processing can be used for OFDM demodulation, replacing the bank of N_c parallel demodulators of Figure 3.5 with sampling with some sampling rate $f_s = 1/T_s$, followed by a size- N DFT/FFT, as illustrated in Figure 3.7.

3.4 CYCLIC-PREFIX INSERTION

As described in Section 3.2, an uncorrupted OFDM signal can be demodulated without any interference between subcarriers. One way to understand this *subcarrier orthogonality* is to recognize that a modulated subcarrier $x_k(t)$ in Eqn (3.1) consists of an integer number of periods of complex exponentials during the demodulator integration interval $T_u = 1/\Delta f$.

However, in case of a time-dispersive channel the orthogonality between the subcarriers will, at least partly, be lost. The reason for this loss of subcarrier orthogonality in the case of a time-dispersive channel is that, in this case, the demodulator correlation interval for one path will overlap with the symbol boundary of a different path, as illustrated in Figure 3.8. Thus, the integration interval will not necessarily correspond to an integer number of periods of complex exponentials of that path as the modulation symbols a_k may differ between consecutive symbol intervals. As a consequence, in the case of a time-dispersive channel there will not only be inter-symbol interference within a subcarrier, but also interference between subcarriers.

Another way to explain the interference between subcarriers in the case of a time-dispersive channel is to have in mind that time dispersion on the radio channel is equivalent to a frequency-selective channel frequency response. As clarified in Section 3.2, orthogonality between OFDM subcarriers is not simply due to frequency-domain separation but due to the *specific* frequency-domain structure of each subcarrier. Even if the frequency-domain channel is constant over a bandwidth corresponding to the main lobe of an OFDM subcarrier and only the subcarrier side lobes are corrupted due to the radio-channel frequency selectivity, the orthogonality between subcarriers will be lost with inter-subcarrier interference as a consequence. Due to the relatively large side lobes of each OFDM subcarrier, already a relatively limited amount of time dispersion or, equivalently, a relatively modest radio-channel frequency selectivity may cause substantial interference between subcarriers.

To deal with this problem and to make an OFDM signal truly insensitive to time dispersion on the radio channel, so-called *cyclic-prefix insertion* is typically used in OFDM transmission. As illustrated

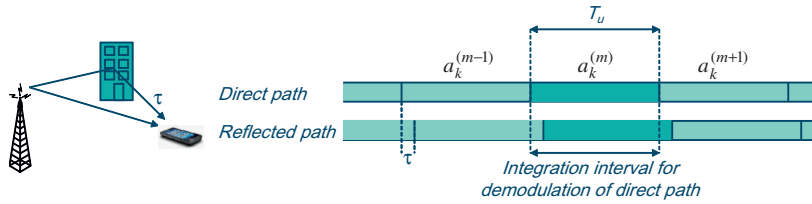


FIGURE 3.8

Time dispersion and corresponding received-signal timing.

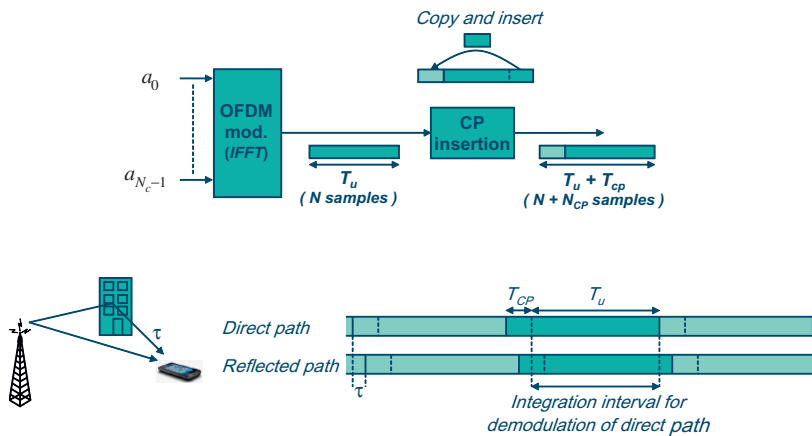


FIGURE 3.9

Cyclic-prefix insertion.

in Figure 3.9, cyclic-prefix insertion implies that the last part of the OFDM symbol is copied and inserted at the beginning of the OFDM symbol. Cyclic-prefix insertion thus increases the length of the OFDM symbol from T_u to $T_u + T_{CP}$, where T_{CP} is the length of the cyclic prefix, with a corresponding reduction in the OFDM symbol rate as a consequence. As illustrated in the lower part of Figure 3.9, if the correlation at the receiver side is still only carried out over a time interval $T_u = 1/\Delta f$, subcarrier orthogonality will then also be preserved in the case of a time-dispersive channel, as long as the span of the time dispersion is shorter than the cyclic-prefix length.

In practice, cyclic-prefix insertion is carried out on the time-discrete output of the transmitter IFFT. Cyclic-prefix insertion then implies that the last N_{CP} samples of the IFFT output block of length N are copied and inserted at the beginning of the block, increasing the block length from N to $N + N_{CP}$. At the receiver side, the corresponding samples are discarded before OFDM demodulation by means of, for example, DFT/FFT processing.

Cyclic-prefix insertion is beneficial in the sense that it makes an OFDM signal insensitive to time dispersion as long as the span of the time dispersion does not exceed the length of the cyclic prefix. The drawback of cyclic-prefix insertion is that only a fraction $T_u/(T_u + T_{CP})$ of the received signal

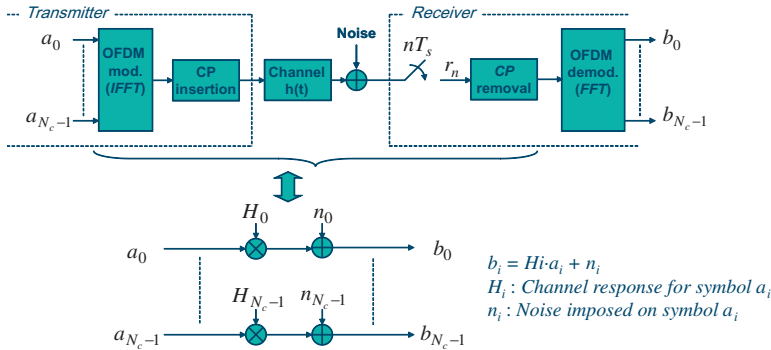


FIGURE 3.10

Frequency-domain model of OFDM transmission/reception.

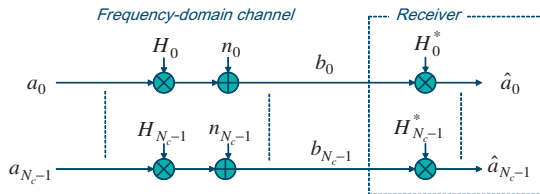
power is actually utilized by the OFDM demodulator, implying a corresponding power loss in the demodulation. In addition to this power loss, cyclic-prefix insertion also implies a corresponding loss in terms of bandwidth as the OFDM symbol rate is reduced without a corresponding reduction in the overall signal bandwidth.

One way to reduce the relative overhead due to cyclic-prefix insertion is to reduce the subcarrier spacing Δf , with a corresponding increase in the symbol time T_u as a consequence. However, this will increase the sensitivity of the OFDM transmission to fast channel variations – that is, high Doppler spread – as well as different types of frequency errors (see further Section 3.8).

It is also important to understand that the cyclic prefix does not necessarily have to cover the entire length of the channel time dispersion. In general, there is a trade-off between the power loss due to the cyclic prefix and the signal corruption (inter-symbol and inter-subcarrier interference) due to residual time dispersion not covered by the cyclic prefix and, at a certain point, further reduction of the signal corruption due to further increase of the cyclic-prefix length will not justify the corresponding additional power loss. This also means that, although the amount of time dispersion typically increases with the cell size, beyond a certain cell size there is often no reason to increase the cyclic prefix further as the corresponding power loss due to a further increase of the cyclic prefix would have a larger negative impact, compared to the signal corruption due to the residual time dispersion not covered by the cyclic prefix [16].

3.5 FREQUENCY-DOMAIN MODEL OF OFDM TRANSMISSION

Assuming a sufficiently large cyclic prefix, the linear convolution of a time-dispersive radio channel will appear as a circular convolution during the demodulator integration interval T_u . The combination of OFDM modulation (IFFT processing), a time-dispersive radio channel, and OFDM demodulation (FFT processing) can then be seen as a *frequency-domain* channel as illustrated in Figure 3.10, where

**FIGURE 3.11**

Frequency-domain model of OFDM transmission/reception with “one-tap equalization” at the receiver.

the frequency-domain channel taps $H_0, \dots, H_{N_c-1}$ can be directly derived from the channel impulse response.

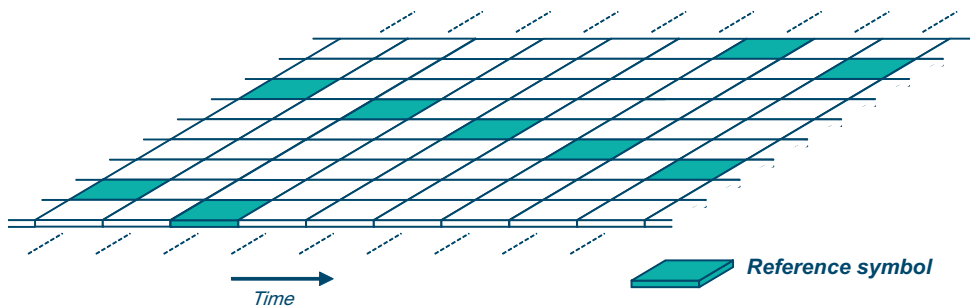
The demodulator output b_k in Figure 3.10 is the transmitted modulation symbol a_k scaled and phase rotated by the complex frequency-domain channel tap H_k and impaired by noise n_k . To properly recover the transmitted symbol for further processing, for example data demodulation and channel decoding, the receiver should multiply b_k by the complex conjugate of H_k , as illustrated in Figure 3.11. This is often expressed as a *one-tap equalizer* being applied to each received subcarrier.

3.6 CHANNEL ESTIMATION AND REFERENCE SYMBOLS

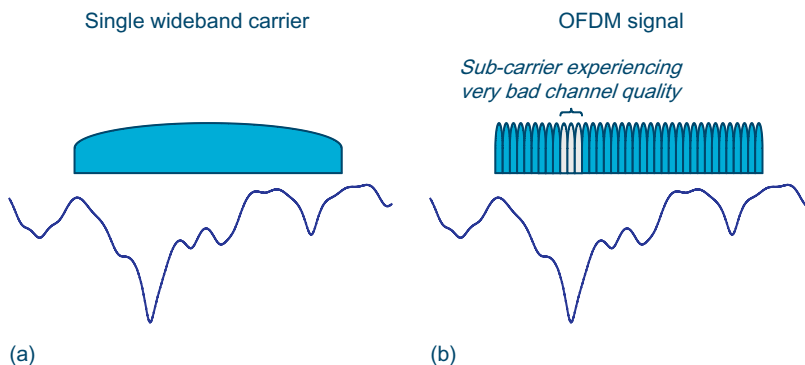
As described above, to demodulate the transmitted modulation symbol a_k and allow for proper decoding of the transmitted information at the receiver side, scaling with the complex conjugate of the frequency-domain channel tap H_k should be applied after OFDM demodulation (*FFT* processing) (see Figure 3.11). To be able to do this, the receiver obviously needs an estimate of the frequency-domain channel taps $H_0, \dots, H_{N_c-1}$.

The frequency-domain channel taps can be estimated indirectly by first estimating the channel impulse response and, from that, calculating an estimate of H_k . However, a more straightforward approach is to estimate the frequency-domain channel taps directly. This can be done by inserting known *reference symbols*, sometimes also referred to as *pilot symbols*, at regular intervals within the OFDM time–frequency grid, as illustrated in Figure 3.12. Using knowledge about the reference symbols, the receiver can estimate the frequency-domain channel around the location of the reference symbol. The reference symbols should also have a sufficiently high density in both the time and the frequency domains to be able to provide estimates for the entire time–frequency grid in the case of radio channels subject to high frequency and/or time selectivity.

Different more or less advanced algorithms can be used for the channel estimation, ranging from simple averaging in combination with linear interpolation to Minimum-Mean-Square-Error (MMSE) estimation relying on more detailed knowledge of the channel time/frequency-domain characteristics. Readers are referred to, for example, [17] for a more in-depth discussion on channel estimation for OFDM.

**FIGURE 3.12**

Time–frequency grid with known reference symbols.

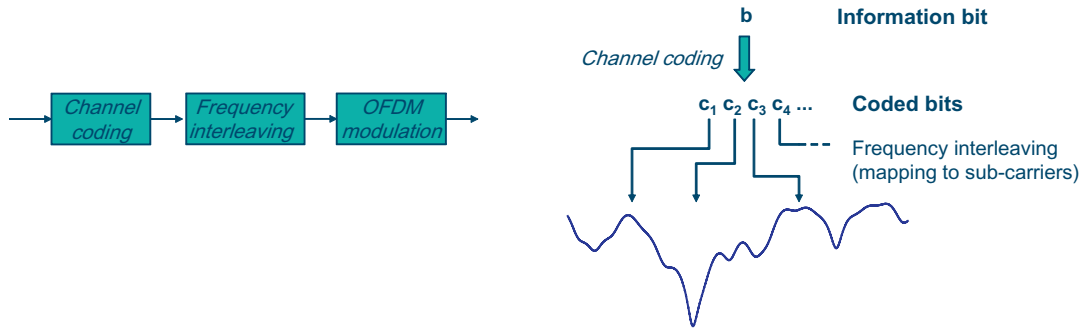
**FIGURE 3.13**

(a) Transmission of single wideband carrier. (b) OFDM transmission over a frequency-selective channel.

3.7 FREQUENCY DIVERSITY WITH OFDM: IMPORTANCE OF CHANNEL CODING

As discussed in Section 2.3 in the previous chapter, a radio channel is always subject to some degree of frequency selectivity, implying that the channel quality will vary in the frequency domain. In the case of a single wideband carrier, such as a WCDMA carrier, each modulation symbol is transmitted over the entire signal bandwidth. Thus, in the case of the transmission of a single wideband carrier over a highly frequency-selective channel (see Figure 3.13a), each modulation symbol will be transmitted both over frequency bands with relatively good quality (relatively high signal strength) and frequency bands with low quality (low signal strength). Such transmission of information over multiple frequency bands with different instantaneous channel quality is also referred to as *frequency diversity*.

On the other hand, in the case of OFDM transmission each modulation symbol is mainly confined to a relatively narrow bandwidth. Thus, for OFDM transmission over a frequency-selective channel, certain modulation symbols may be fully confined to a frequency band with very low instantaneous

**FIGURE 3.14**

Channel coding in combination with frequency-domain interleaving to provide frequency diversity in the case of OFDM transmission.

signal strength, as illustrated in Figure 3.13b. Thus, the individual modulation symbols will typically not experience any substantial frequency diversity even if the channel is highly frequency selective over the overall OFDM transmission bandwidth. As a consequence, the basic error-rate performance of OFDM transmission over a frequency-selective channel is relatively poor and especially much worse than the basic error rate in the case of a single wideband carrier.

However, in practice channel coding is used in most cases of digital communication and especially in mobile communication. Channel coding implies that each bit of information to be transmitted is spread over several, often very many, code bits. If these coded bits are then, via modulation symbols, mapped to a set of OFDM subcarriers that are well distributed over the overall transmission bandwidth of the OFDM signal, as illustrated in Figure 3.14, each information bit will experience frequency diversity in the case of transmission over a radio channel that is frequency selective over the transmission bandwidth, despite the fact that the subcarriers, and thus also the code bits, will not experience any frequency diversity. Distributing the code bits in the frequency domain, as illustrated in Figure 3.14, is sometimes referred to as *frequency interleaving*. This is similar to the use of time-domain interleaving to benefit from channel coding in the case of fading that varies in time.

Thus, in contrast to the transmission of a single wideband carrier, channel coding (combined with frequency interleaving) is an essential component in order for OFDM transmission to be able to benefit from frequency diversity on a frequency-selective channel. As channel coding is typically used in most cases of mobile communication this is not a very serious drawback, especially taking into account that a significant part of the available frequency diversity can be captured already with a relatively high code rate.

3.8 SELECTION OF BASIC OFDM PARAMETERS

If OFDM is to be used as the transmission scheme in a mobile-communication system, the following basic OFDM parameters need to be decided upon:

- The subcarrier spacing Δf .
- The number of subcarriers N_c , which, together with the subcarrier spacing, determines the overall transmission bandwidth of the OFDM signal.

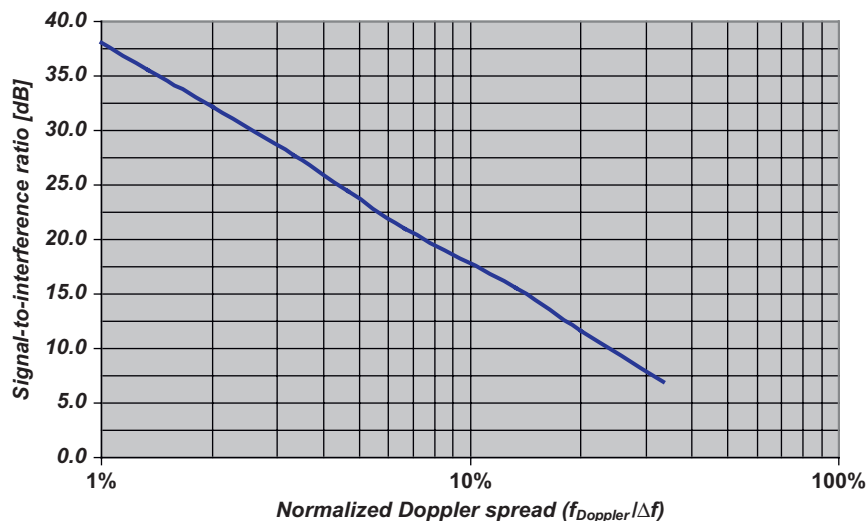


FIGURE 3.15

Subcarrier interference as a function of the normalized Doppler spread $f_{\text{Doppler}}/\Delta f$.

- The cyclic-prefix length T_{CP} . Together with the subcarrier spacing $\Delta f = 1/T_u$, the cyclic-prefix length determines the overall OFDM symbol time $T = T_{\text{CP}} + T_u$ or, equivalently, the OFDM symbol rate.

3.8.1 OFDM Subcarrier Spacing

There are two factors that constrain the selection of the OFDM subcarrier spacing:

- The OFDM subcarrier spacing should be as small as possible (T_u as large as possible) to minimize the relative cyclic-prefix overhead $T_{\text{CP}}/(T_u + T_{\text{CP}})$ – see further Section 3.8.3.
- A too small subcarrier spacing increases the sensitivity of the OFDM transmission to Doppler spread and different kinds of frequency inaccuracies.

A requirement for the OFDM subcarrier orthogonality (3.2) to hold at the receiver side – that is, after the transmitted signal has propagated over the radio channel – is that the instantaneous channel does not vary noticeably during the demodulator correlation interval T_u (see Figure 3.5). In the case of such channel variations, for example due to very high Doppler spread, the orthogonality between subcarriers will be lost, with inter-subcarrier interference as a consequence. Figure 3.15 illustrates the subcarrier signal-to-interference ratio due to inter-subcarrier interference between two neighboring subcarriers, as a function of the normalized Doppler spread. When considering Figure 3.15, it should be borne in mind that a subcarrier will be subject to interference from multiple subcarriers on both sides⁴ – that is, the overall inter-subcarrier interference from all subcarriers will be higher than what is illustrated in Figure 3.15.

⁴Except for the subcarrier at the edge of the spectrum.

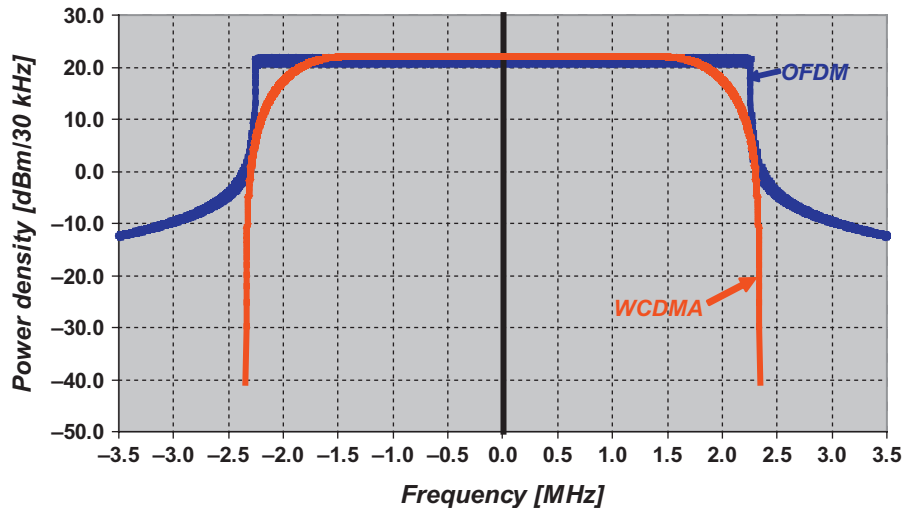


FIGURE 3.16

Spectrum of a basic 5 MHz OFDM signal compared with WCDMA spectrum.

In practice, the amount of inter-subcarrier interference that can be accepted very much depends on the service to be provided and to what extent the received signal is anyway corrupted due to noise and other impairments. As an example, on the cell border of large cells the signal-to-noise/interference ratio will anyway be relatively low, with relatively low achievable data rates as a consequence. A small amount of additional inter-subcarrier interference, for example due to Doppler spread, may then be more or less negligible. At the same time, in high signal-to-noise/interference scenarios, for example in small cells with low traffic or close to the base station, where high data rates are to be provided, the same amount of inter-subcarrier interference may have a much more negative impact.

It should also be noted that, in addition to Doppler spread, inter-subcarrier interference will also be due to different transmitter and receiver inaccuracies, such as frequency errors and phase noise.

3.8.2 Number of Subcarriers

Once the subcarrier spacing has been selected based on environment, expected Doppler spread and time dispersion, etc., the number of subcarriers can be determined based on the amount of spectrum available and the acceptable out-of-band emissions.

The basic bandwidth of an OFDM signal equals $N_c \cdot \Delta f$ – that is, the number of subcarriers multiplied by the subcarrier spacing. However, as can be seen in Figure 3.16, the spectrum of a basic OFDM signal falls off very slowly outside the basic OFDM bandwidth and especially much slower than for a WCDMA signal. The reason for the large *out-of-band emission* of a basic OFDM signal is the use of rectangular pulse shaping (Figure 3.1), leading to per-subcarrier side lobes that fall off relatively slowly. However, in practice, straightforward filtering or *time-domain windowing* [18] will be used to suppress a main part of the OFDM out-of-band emissions. Thus, in practice, typically of the

order of 10% guard-band is needed for an OFDM signal implying, as an example, that in a spectrum allocation of 5 MHz, the basic OFDM bandwidth $N_c \cdot \Delta f$ could be of the order of 4.5 MHz. Assuming, for example, a subcarrier spacing of 15 kHz as selected for LTE, this corresponds to approximately 300 subcarriers in 5 MHz.

3.8.3 Cyclic-Prefix Length

In principle, the cyclic-prefix length T_{CP} should cover the maximum length of the time dispersion expected to be experienced. However, as already discussed, increasing the length of the cyclic prefix, without a corresponding reduction in the subcarrier spacing Δf , implies an additional overhead in terms of power as well as bandwidth. In particular, the power loss implies that, as the cell size grows and the system performance becomes more power limited, there is a trade-off between the loss in power due to the cyclic prefix and the signal corruption due to time dispersion not covered by the cyclic prefix. As already mentioned, this implies that, although the amount of time dispersion typically increases with the cell size, beyond a certain cell size there is often no reason to increase the cyclic prefix further as the corresponding power loss would have a larger negative impact, compared to the signal corruption due to the residual time dispersion not covered by the cyclic prefix [16].

One situation where a longer cyclic prefix may be needed is in the case of multi-cell transmission using SFN (Single-Frequency Network), as further discussed in Section 3.11.

Thus, to be able to optimize performance to different environments, some OFDM-based systems support multiple cyclic-prefix lengths. The different cyclic-prefix lengths can then be used in different transmission scenarios:

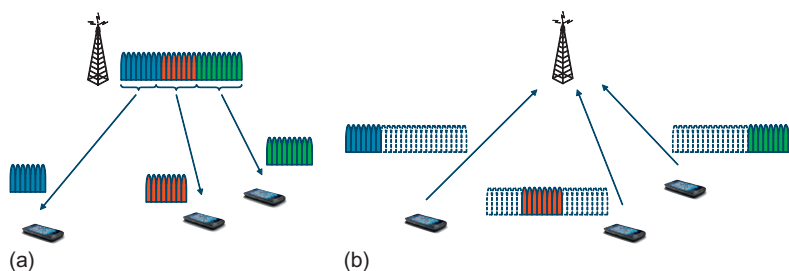
- Shorter cyclic prefix in small-cell environments to minimize the cyclic-prefix overhead.
- Longer cyclic prefix in environments with extreme time dispersion and especially in the case of SFN operation.

3.9 VARIATIONS IN INSTANTANEOUS TRANSMISSION POWER

According to Section 2.3.1, one of the drawbacks of multi-carrier transmission is the corresponding large variations in the instantaneous transmit power, implying a reduced power-amplifier efficiency and higher mobile-terminal power consumption; alternatively, that the power-amplifier output power has to be reduced with a reduced range as a consequence. Being a kind of multi-carrier transmission scheme, OFDM is subject to the same drawback.

However, a large number of different methods have been proposed to reduce the large power peaks of an OFDM signal:

- In the case of *tone reservation* [19], a subset of the OFDM subcarriers are not used for data transmission. Instead, these subcarriers are modulated in such a way that the largest peaks of the overall OFDM signal are suppressed, allowing for a reduced power-amplifier back-off. One drawback of tone reservation is the bandwidth loss due to the fact that a number of subcarriers are not available for actual data transmission. The calculation of what modulation to apply to the reserved tones can also be of relatively high complexity.
- In the case of *prefiltering* or *precoding*, linear processing is applied to the sequence of modulation symbols before OFDM modulation. DFTS-spread OFDM (DFTS-OFDM) as described in the next chapter and which is used for the LTE uplink, can be seen as one kind of prefiltering.

**FIGURE 3.17**

OFDM as a user-multiplexing/multiple-access scheme: (a) downlink; (b) uplink.

- In the case of *selective scrambling* [20], the coded-bit sequence to be transmitted is scrambled with a number of different scrambling codes. Each scrambled sequence is then OFDM modulated and the signal with the lowest peak power is selected for transmission. After OFDM demodulation at the receiver side, descrambling and subsequent decoding is carried out for all the possible scrambling sequences. Only the decoding carried out for the scrambling code actually used for the transmission will provide a correct decoding result. A drawback of selective scrambling is an increased receiver complexity as multiple decodings need to be carried out in parallel.

Readers are referred to the references above for a more in-depth discussion on different peak-reduction schemes.

3.10 OFDM AS A USER-MULTIPLEXING AND MULTIPLE-ACCESS SCHEME

The discussion has, until now, implicitly assumed that all OFDM subcarriers are transmitted from the same transmitter to a certain receiver, that is:

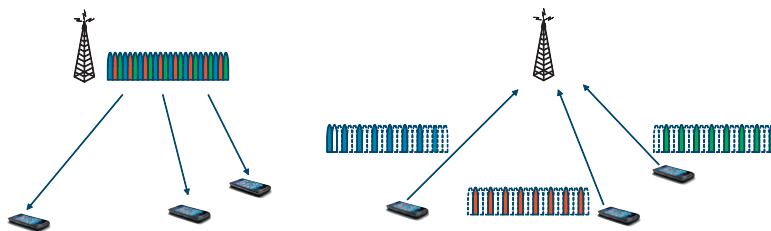
- downlink transmission of all subcarriers to a *single* terminal;
- uplink transmission of all subcarriers from a *single* terminal.

However, OFDM can also be used as a *user-multiplexing* or *multiple-access scheme*, allowing for simultaneous frequency-separated transmissions to/from multiple terminals (see Figure 3.17).

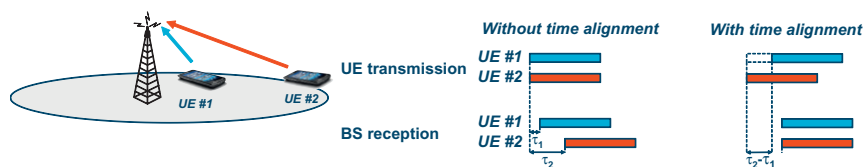
In the downlink direction, OFDM as a user-multiplexing scheme implies that, in each OFDM symbol interval, different subsets of the overall set of available subcarriers are used for transmission to *different* terminals (see Figure 3.17a).

Similarly, in the uplink direction, OFDM as a *user-multiplexing* or *multiple-access scheme* implies that, in each OFDM symbol interval, different subsets of the overall set of subcarriers are used for data transmission from *different* terminals (see Figure 3.17b). In this case, the term Orthogonal Frequency-Division Multiple Access or OFDMA is also often used.⁵

⁵The term OFDMA is sometimes also used to denote the use of OFDM to multiplex multiple users in the downlink, as illustrated in Figure 3.17a.

**FIGURE 3.18**

Distributed user multiplexing.

**FIGURE 3.19**

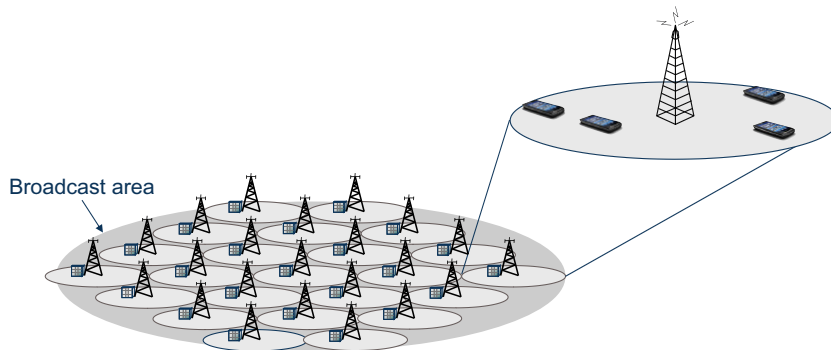
Uplink transmission-timing control.

Figure 3.17 assumes that *consecutive* subcarriers are used for transmission to/from the same terminal. However, distributing the subcarriers to/from a terminal in the frequency domain is also possible, as illustrated in Figure 3.18. The benefit of such *distributed* user multiplexing or *distributed* multiple access is a possibility for additional frequency diversity as each transmission is spread over a wider bandwidth.

In the case when OFDMA is used as an uplink multiple-access scheme – that is, in the case of frequency multiplexing of OFDM signals from multiple terminals – it is critical that the transmissions from the different terminals arrive approximately time aligned at the base station. More specifically, the transmissions from the different terminals should arrive at the base station with a timing misalignment less than the length of the cyclic prefix to preserve orthogonality between subcarriers received from different terminals and thus avoid inter-user interference.

Due to the differences in distance to the base station for different terminals and the corresponding differences in the propagation time (which may far exceed the length of the cyclic prefix), it is therefore necessary to control the uplink transmission timing of each terminal (see Figure 3.19). Such *transmission-timing control* should adjust the transmit timing of each terminal to ensure that uplink transmissions arrive approximately time aligned at the base station. As the propagation time changes as the terminal is moving within the cell, the transmission-timing control should be an active process, continuously adjusting the exact transmit timing of each terminal.

Furthermore, even in the case of perfect transmission-timing control, there will always be some interference between subcarriers, for example due to frequency errors. Typically this interference is relatively low in the case of reasonable frequency errors, Doppler spread, etc. (see Section 3.8). However, this assumes that the different subcarriers are received with at least approximately the same

**FIGURE 3.20**

Broadcast scenario.

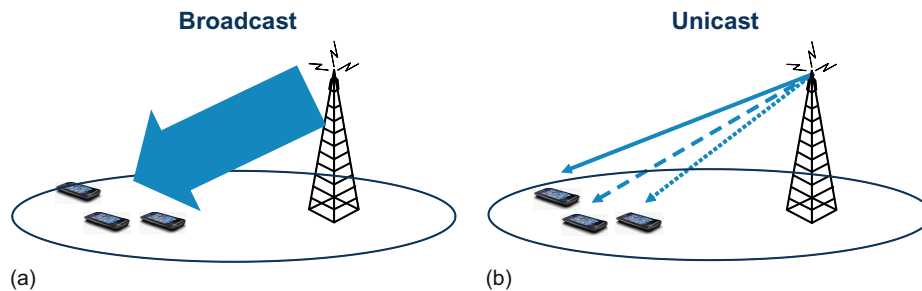
power. In the uplink, the propagation distance and thus the path loss of the different mobile-terminal transmissions may differ significantly. If two terminals are transmitting with the same power, the received-signal strengths may thus differ significantly, implying a potentially significant interference from the stronger signal to the weaker signal unless the subcarrier orthogonality is perfectly retained. To avoid this, at least some degree of *uplink transmission-power control* may need to be applied in case of uplink OFDMA, reducing the transmission power of user terminals close to the base station and ensuring that all received signals will be of approximately the same power.

3.11 MULTI-CELL BROADCAST/MULTICAST TRANSMISSION AND OFDM

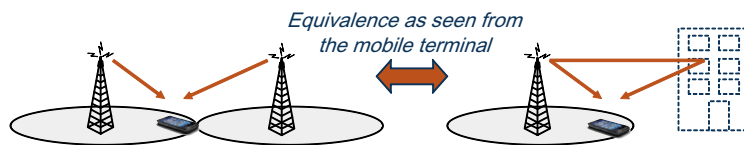
The provision of broadcast/multicast services in a mobile-communication system implies that the same information is to be *simultaneously* provided to multiple terminals, often dispersed over a large area corresponding to a large number of cells, as shown in Figure 3.20. The broadcast/multicast information may be a TV news clip, information about the local weather conditions, stock-market information, or any other kind of information that, at a given time, may be of interest to a large number of people.

When the same information is to be provided to multiple terminals within a cell it is often beneficial to provide this information as a single “broadcast” radio transmission covering the entire cell and simultaneously being received by all relevant terminals (Figure 3.21a), rather than providing the information by means of individual transmissions to each terminal (unicast transmission; see Figure 3.21b).

As a broadcast transmission according to Figure 3.21 also has to be dimensioned to reach the worst-case terminals, including terminals at the cell border, it will be relative costly in terms of the recourses (base-station transmit power) needed to provide a certain broadcast-service data rate. Alternatively, taking into account the limited signal-to-noise ratio that can be achieved at, for example, the cell edge, the achievable broadcast data rates may be relatively limited, especially in the case of large cells. One way to increase the broadcast data rates would then be to reduce the cell size, thereby increasing the cell-edge receive power. However, this will increase the number of cells to cover a certain area and is thus obviously negative from a cost-of-deployment point of view.

**FIGURE 3.21**

Broadcast vs. unicast transmission. (a) Broadcast. (b) Unicast.

**FIGURE 3.22**

Equivalence between simulcast transmission and multi-path propagation.

However, as discussed above, the provision of broadcast/multicast services in a mobile-communication network typically implies that identical information is to be provided over a large number of cells. In such a case, the resources (downlink transmit power) needed to provide a certain broadcast data rate can be considerably reduced if terminals at the cell edge can utilize the received power from broadcast transmissions from multiple cells when detecting/decoding the broadcast data.

One way to mitigate this and further improve the provision of broadcast/multicast services in a mobile-communication network is to ensure that the broadcast transmissions from different cells *are truly identical and transmitted mutually time aligned*. In this case, the transmissions received from multiple cells will, as seen from the terminal, appear as a single transmission subject to severe multi-path propagation, as illustrated in Figure 3.22. The transmission of identical time-aligned signals from multiple cells, especially in the case of provision of broadcast/multicast services, is sometimes referred to as Single-Frequency Network (SFN) operation [21].

In the case of such identical time-aligned transmissions from multiple cells, the “inter-cell interference” due to transmissions in neighboring cells will, from a terminal point of view, be replaced by signal corruption due to time dispersion. If the broadcast transmission is based on OFDM with a cyclic prefix that covers the main part of this “time dispersion”, the achievable broadcast data rates are thus only limited by noise, implying that, especially in smaller cells, very high broadcast data rates can be achieved. Furthermore, the OFDM receiver does not need to explicitly identify the cells to be soft combined. Rather, all transmissions that fall within the cyclic prefix will “automatically” be captured by the receiver.

Wider-Band “Single-Carrier” Transmission

The previous chapters discussed multi-carrier transmission in general (Section 2.3.1) and OFDM transmission in particular (Chapter 3) as means to allow for very high overall transmission bandwidth while still being robust to signal corruption due to radio-channel frequency selectivity.

However, as discussed in the previous chapter, a drawback of OFDM modulation, as well as any kind of multi-carrier transmission, is the large variations in the instantaneous power of the transmitted signal. Such power variations imply a reduced power-amplifier efficiency and higher power-amplifier cost. This is especially critical for the uplink, due to the high importance of low mobile-terminal power consumption and cost.

As discussed in Chapter 3, several methods have been proposed on how to reduce the large power variations of an OFDM signal. However, most of these methods have limitations in terms of to what extent the power variations can be reduced. Furthermore, most of the methods also imply a significant computational complexity and/or a reduced link performance.

Thus, there is an interest to consider also wider-band *single-carrier* transmission as an alternative to multi-carrier transmission, especially for the uplink – that is, for mobile-terminal transmissions. It is then necessary to consider what can be done to handle the corruption to the signal waveform that will occur in most mobile-communication environments due to radio-channel frequency selectivity.

4.1 EQUALIZATION AGAINST RADIO-CHANNEL FREQUENCY SELECTIVITY

Historically, the main method to handle signal corruption due to radio-channel frequency selectivity has been to apply different forms of *equalization* [12] at the receiver side. The aim of equalization is to, by different means, compensate for the channel frequency selectivity and thus, at least to some extent, restore the original signal shape.

4.1.1 Time-Domain Linear Equalization

The most basic approach to equalization is the time-domain linear equalizer, consisting of a linear filter with an impulse response $w(\tau)$ applied to the received signal (see Figure 4.1).

By selecting different filter impulse responses, different receiver/equalizer strategies can be implemented. As an example, in DS-CDMA-based systems a so-called RAKE receiver structure has historically often been used. The RAKE receiver is simply the receiver structure of Figure 4.1 where the filter impulse response has been selected to provide *channel-matched filtering*:

$$w(\tau) = h^*(-\tau). \quad (4.1)$$

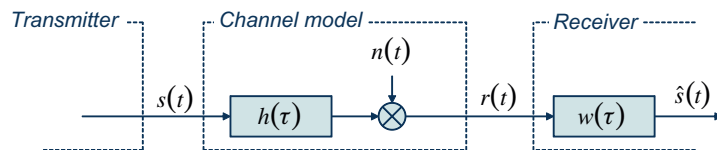


FIGURE 4.1

General time-domain linear equalization.

That is, the filter response has been selected as the complex conjugate of the time-reversed channel impulse response. This is also often referred to as a *Maximum-Ratio Combining* (MRC) filter setting [12].

Selecting the receiver filter according to the MRC criterion – that is, as a channel-matched filter – maximizes the post-filter signal-to-noise ratio (thus the term maximum-ratio combining). However, MRC-based filtering does not provide any compensation for any radio-channel frequency selectivity – that is, no equalization. Thus, MRC-based receiver filtering is appropriate when the received signal is mainly impaired by noise or interference from other transmissions but not when a main part of the overall signal corruption is due to the radio-channel frequency selectivity.

Another alternative is to select the receiver filter to fully compensate for the radio-channel frequency selectivity. This can be achieved by selecting the receiver-filter impulse response to fulfill the relation:

$$h(\tau) \otimes w(\tau) = 1, \quad (4.2)$$

where “ $\otimes$ ” denotes linear convolution. This selecting of the filter setting, also known as *Zero-Forcing* (ZF) equalization [12], provides full compensation for any radio-channel frequency selectivity (complete equalization) and thus full suppression of any related signal corruption. However, zero-forcing equalization may lead to a large, potentially very large, increase in the noise level after equalization and thus to an overall degradation in the link performance. This will be the case especially when the channel has large variations in its frequency response.

A third and, in most cases, better alternative is to select a filter setting that provides a trade-off between signal corruption due to radio-channel frequency selectivity, and noise/interference. This can, for example, be done by selecting the filter to minimize the mean-square error between the equalizer output and the transmitted signal – that is, to minimize:

$$\varepsilon = E \left\{ | \hat{s}(t) - s(t) |^2 \right\}. \quad (4.3)$$

This is also referred to as a *Minimum Mean-Square Error* (MMSE) equalizer setting [12].

In practice, the linear equalizer has most often been implemented as a *time-discrete FIR filter* [22] with L filter taps applied to the sampled received signal, as illustrated in Figure 4.2. In general, the complexity of such a time-discrete equalizer grows relatively rapidly with the bandwidth of the signal to be equalized:

- A more wideband signal is subject to relatively more radio-channel frequency selectivity or, equivalently, relatively more time dispersion. This implies that the equalizer needs to have a larger

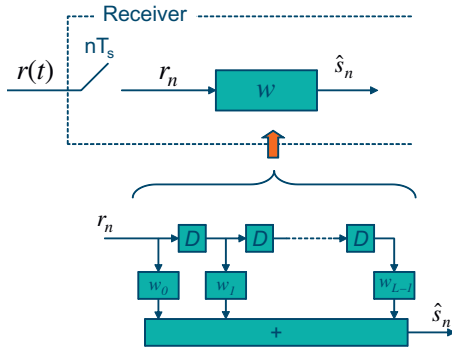


FIGURE 4.2

Linear equalization implemented as a time-discrete FIR filter.

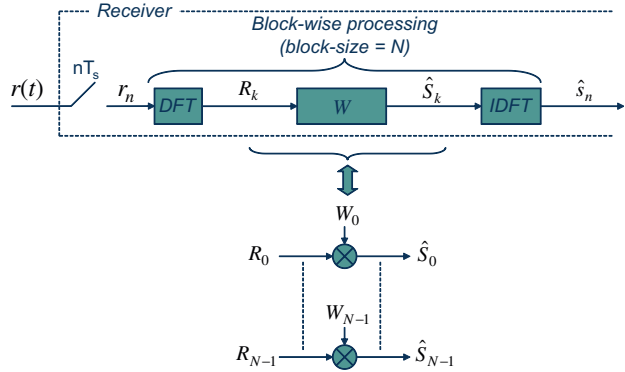


FIGURE 4.3

Frequency-domain linear equalization.

span (larger length L – that is, more filter taps) to be able to properly compensate for the channel frequency selectivity.

- A more wideband signal leads to a correspondingly higher sampling rate for the received signal. Thus, also the receiver-filter processing needs to be carried out with a correspondingly higher rate.

It can be shown [23] that the time-discrete MMSE equalizer setting $\bar{w} = [w_0, w_1, \dots, w_{L-1}]^H$ is given by the expression:

$$\bar{w} = R^{-1} \bar{p}. \quad (4.4)$$

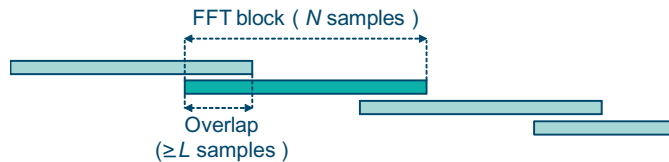
In this expression, R is the *channel-output auto-correlation matrix* of size $L \times L$, which depends on the channel impulse response as well as on the noise level, and $\bar{p}$ is the *channel-output/channel-input cross-correlation vector* of size $L \times 1$ that depends on the channel impulse response.

In particular, in the case of a large equalizer span (large L), the time-domain MMSE equalizer may be of relatively high complexity:

- The equalization itself (the actual filtering) may be of relatively high complexity according to the above.
- Calculation of the MMSE equalizer setting, especially the calculation of the inverse of the size $L \times L$ correlation matrix R , may be of relatively high complexity.

4.1.2 Frequency-Domain Equalization

A possible way to reduce the complexity of linear equalization is to carry out the equalization in the frequency domain [24], as illustrated in Figure 4.3. In such *frequency-domain linear equalization*, the equalization is carried out block-wise with block size N . The sampled received signal is first transformed into the frequency domain by means of a size- N DFT. The equalization is then carried out as

**FIGURE 4.4**

Overlap-and-discard processing.

frequency-domain filtering, with the frequency-domain filter taps $W_0, \dots, W_{N-1}$, for example, being the DFT of the corresponding time-domain filter taps $w_0, \dots, w_{L-1}$ of Figure 4.2. Finally, the equalized frequency-domain signal is transformed back to the time domain by means of a size- N inverse DFT. The block size- N should preferably be selected as $N = 2^n$ for some integer n to allow for computational-efficient radix-2 FFT/IFFT implementation of the DFT/IDFT processing.

For each processing block of size N , the frequency-domain equalization basically consists of:

- A size- N DFT/FFT.
- N complex multiplications (the frequency-domain filter).
- A size- N inverse DFT/FFT.

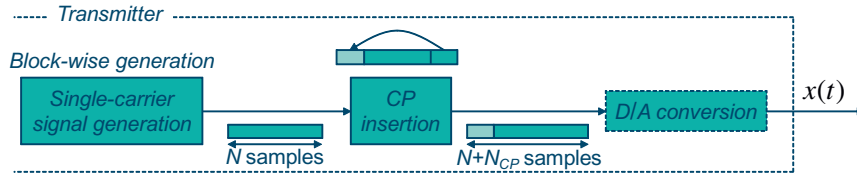
Especially in the case of channels with extensive frequency selectivity, implying the need for a large span of a time-domain equalizer (large equalizer length L), equalization in the frequency domain according to Figure 4.3 can be of significantly less complexity, compared to time-domain equalization illustrated in Figure 4.2.

However, there are two issues with frequency-domain equalization:

- The time-domain filtering of Figure 4.2 implements a time-discrete *linear convolution*. In contrast, frequency-domain filtering according to Figure 4.3 corresponds to *circular convolution* in the time domain. Assuming a time-domain equalizer of length L , this implies that the first $L - 1$ samples at the output of the frequency-domain equalizer *will not* be identical to the corresponding output of the time-domain equalizer.
- The frequency-domain filter taps $W_0, \dots, W_{N-1}$ can be determined by first determining the pulse response of the corresponding time-domain filter and then transforming this filter into the frequency domain by means of a DFT. However, as mentioned above, determining the MMSE time-domain filter may be relatively complex in the case of a large equalizer length L .

One way to address the first issue is to apply an *overlap* in the block-wise processing of the frequency-domain equalizer as outlined in Figure 4.4, where the overlap should be at least $L - 1$ samples. With such an overlap, the first $L - 1$ “incorrect” samples at the output of the frequency-domain equalizer can be *discarded* as the corresponding samples are also (correctly) provided as the last part of the previously received/equalized block. The drawback with this kind of “overlap-and-discard” processing is a computational overhead – that is, somewhat higher receiver complexity.

An alternative approach that addresses both of the above issues is to apply *cyclic-prefix insertion* at the transmitter side (see Figure 4.5). Similar to OFDM, cyclic-prefix insertion in the case of single-carrier transmission implies that a cyclic prefix of length N_{CP} samples is inserted block-wise at the

**FIGURE 4.5**

Cyclic-prefix insertion in single-carrier transmission.

transmitter side. The transmitter-side block size should be the same as the block size N used for the receiver-side frequency-domain equalization.

With the introduction of a cyclic prefix, the channel will, from a receiver point of view, appear as a circular convolution over a receiver processing block of size N . Thus, there is no need for any receiver overlap-and-discard processing. Furthermore, the frequency-domain filter taps can now be calculated directly from an estimate of the sampled channel frequency response without first determining the time-domain equalizer setting. As an example, in the case of an MMSE equalizer the frequency-domain filter taps can be calculated according to:

$$W_k = \frac{H_k^*}{|H_k|^2 + N_0}, \quad (4.5)$$

where N_0 is the noise power and H_k is the sampled channel frequency response. For large equalizer lengths, this calculation is of much lower complexity compared to the time-domain calculation discussed in the previous section.

The drawback of cyclic-prefix insertion in the case of single-carrier transmission is the same as for OFDM – that is, it implies an overhead in terms of both power and bandwidth. One method to reduce the relative cyclic-prefix overhead is to increase the block size N of the frequency-domain equalizer. However, for the block-wise equalization to be accurate, the channel needs to be approximately constant over a time span corresponding to the size of the processing block. This constraint provides an upper limit on the block size N that depends on the rate of the channel variations. Note that this is similar to the constraint on the OFDM subcarrier spacing $\Delta f = 1/T_u$ depending on the rate of the channel variations, as discussed in Chapter 3.

4.1.3 Other Equalizer Strategies

The previous sections discussed different approaches to linear equalization as a means to counteract signal corruption of a wideband signal due to radio-channel frequency selectivity. However, there are also other approaches to equalization:

- *Decision-Feedback Equalization* (DFE) [12] implies that previously detected symbols are fed back and used to cancel the contribution of the corresponding transmitted symbols to the overall signal corruption. Such decision feedback is typically used in combination with time-domain linear filtering, where the linear filter transforms the channel response to a shape that is more

suitable for the decision-feedback stage. Decision feedback can also very well be used in combination with frequency-domain linear equalization [24].

- *Maximum-Likelihood* (ML) detection, also known as *Maximum-Likelihood Sequence Estimation* (MLSE) [25], is strictly speaking not an equalization scheme but rather a receiver approach where the impact of radio-channel time dispersion is explicitly taken into account in the receiver-side detection process. Fundamentally, an ML detector uses the entire received signal to decide on the most likely transmitted sequence, taking into account the impact of the time dispersion on the received signal. To implement maximum-likelihood detection the *Viterbi algorithm* [26] is often used.¹ However, although maximum-likelihood detection based on the Viterbi algorithm has been extensively used for 2G mobile communication such as GSM, it can be too complex when applied to much wider transmission bandwidths, leading to both much more extensive channel frequency selectivity and much higher sampling rates.

4.2 UPLINK FDMA WITH FLEXIBLE BANDWIDTH ASSIGNMENT

In practice, there are obviously often multiple terminals within a cell and thus multiple terminals that should share the overall uplink radio resource within the cell by means of the *uplink intra-cell multiple-access scheme*.

In mobile communication based on WCDMA and CDMA2000, uplink transmissions within a cell are mutually *non-orthogonal* and the base-station receiver relies on the processing gain due to channel coding and additional direct-sequence spreading to suppress the intra-cell interference. Although non-orthogonal multiple access can, fundamentally, provide higher capacity compared to orthogonal multiple access, in practice the possibility for mutually orthogonal uplink transmissions within a cell is often beneficial from a system performance point of view.

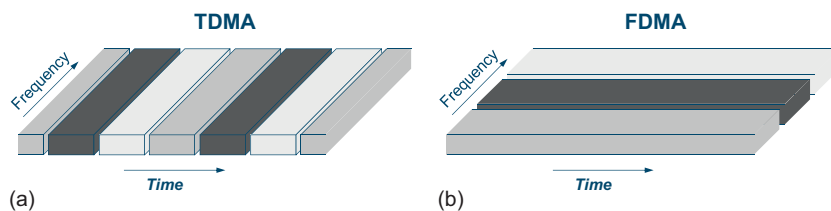
Mutually orthogonal uplink transmissions within a cell can be achieved in the time domain (*Time-Division Multiple Access*, TDMA), as illustrated in Figure 4.6a. TDMA implies that different terminals within a cell transmit in different non-overlapping time intervals. In each such time interval, the full system bandwidth is then assigned for uplink transmission from a single terminal.

Alternatively, mutually orthogonal uplink transmissions within a cell can be achieved in the frequency domain (*Frequency-Division Multiple Access*, FDMA), as illustrated in Figure 4.6b – that is, by having terminals transmit in different frequency bands.

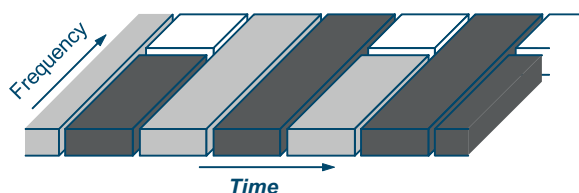
To be able to provide high-rate packet-data transmission, it should be possible to assign the entire system bandwidth for transmission from a single terminal. At the same time, due to the burstiness of most packet-data services, in many cases terminals will have no uplink data to transmit. Thus, for efficient packet-data access, a TDMA component should always be part of the uplink multiple-access scheme.

However, relying on only TDMA to provide orthogonality between uplink transmissions within a cell could be bandwidth inefficient, especially in case of a very wide overall system bandwidth. As discussed in Chapter 2, a wide bandwidth is needed to support high data rates in a power-efficient way. However, the data rates that can be achieved over a radio link are, in many cases, limited by the

¹Thus, the term *Viterbi equalizer* is sometimes used for the ML detector.

**FIGURE 4.6**

Orthogonal multiple access: (a) TDMA; (b) FDMA.

**FIGURE 4.7**

FDMA with flexible bandwidth assignment.

available signal power (power-limited operation) rather than by the available bandwidth. This is especially the case for the uplink, due to the, in general, more limited mobile-terminal transmit power. Allocating the entire system bandwidth to a single terminal could, in such cases, be highly inefficient in terms of bandwidth utilization. As an example, allocating 20 MHz of transmission bandwidth to a terminal in a scenario where the achievable uplink data rate, due to mobile-terminal transmit-power limitations, is anyway limited to, for example, a few 100 kbit/s would obviously imply a very inefficient usage of the overall available bandwidth. In such cases, a smaller transmission bandwidth should be assigned to the terminal and the remaining part of the overall system bandwidth should be used for uplink transmissions from other terminals. Thus, in addition to TDMA, an uplink transmission scheme should preferably allow for orthogonal user multiplexing also in the frequency domain – that is, FDMA.

At the same time, it should be possible to allocate the entire overall transmission bandwidth to a single terminal when the channel conditions are such that the wide bandwidth can be efficiently utilized – that is, when the achievable data rates are not power limited. Thus, an orthogonal uplink transmission scheme should allow for *FDMA with flexible bandwidth assignment*, as illustrated in Figure 4.7.

Flexible bandwidth assignment is straightforward to achieve with an OFDM-based uplink transmission scheme by dynamically allocating different number of subcarriers to different terminals depending on their instantaneous channel conditions. In the next section, it will be discussed how this can also be achieved in the case of low-PAR “single-carrier” transmission, more specifically by means of so-called *DFT-spread OFDM*.

4.3 DFT-SPREAD OFDM

DFT-spread OFDM (DFTS-OFDM) is a transmission scheme that can combine the desired properties discussed in the previous sections, that is:

- Small variations in the instantaneous power of the transmitted signal (“single-carrier” property).
- Possibility for low-complexity high-quality equalization in the frequency domain.
- Possibility for FDMA with flexible bandwidth assignment.

Due to these properties, DFTS-OFDM is, for example, used for uplink data transmission in LTE.

4.3.1 Basic Principles

The basic principle of DFTS-OFDM transmission is illustrated in Figure 4.8. One way to interpret DFTS-OFDM is as normal OFDM with a *DFT-based precoding*. Similar to OFDM modulation, DFTS-OFDM relies on block-based signal generation. In case of DFTS-OFDM, a block of M modulation symbols from some modulation alphabet, for example QPSK or 16QAM, is first applied to a size- M DFT. The output of the DFT is then applied to consecutive inputs (subcarriers) of an OFDM modulator where, in practice, the OFDM modulator will be implemented as a size- N inverse DFT (IDFT) with $N > M$ and where the unused inputs of the IDFT are set to zero. Typically, the IDFT size N is selected as $N = 2^n$ for some integer n to allow for the IDFT to be implemented by means of computationally efficient radix-2 IFFT processing. Also, similar to normal OFDM, a cyclic prefix is preferably inserted for each transmitted block. As discussed in Section 4.1.2, the presence of a cyclic prefix allows for straightforward low-complexity frequency-domain equalization at the receiver side.

If the DFT size M equals the IDFT size N , the cascaded DFT/IDFT processing would obviously completely cancel each other out. However, if M is smaller than N and the remaining inputs to the IDFT are set to zero, the output of the IDFT will be a signal with “single-carrier” properties – that is, a signal with low power variations, and with a bandwidth that depends on M . More specifically, assuming a sampling rate f_s at the output of the IDFT, the nominal bandwidth of the transmitted signal will be $BW = M/N \cdot f_s$. Thus, by varying the block size M the instantaneous bandwidth of the transmitted signal can be varied, allowing for flexible-bandwidth assignment. Furthermore, by shifting the IDFT inputs to which the DFT outputs are mapped, the transmitted signal can be shifted in the frequency domain, as will be further discussed in Section 4.3.3.

To have a high degree of flexibility in the instantaneous bandwidth, given by the DFT size M , it is typically not possible to ensure that M can be expressed as 2^m for some integer m . However, as long as M can be expressed as a product of relatively small prime numbers, the DFT can still be implemented as relatively low-complexity non-radix-2 FFT processing. As an example, a DFT size $M = 144$ can be implemented by means of a combination of radix-2 and radix-3 FFT processing ($144 = 3^2 \cdot 2^4$).

The main benefit of DFTS-OFDM, compared to normal OFDM, is reduced variations in the instantaneous transmission power, implying the possibility for increased power-amplifier efficiency. This benefit of DFTS-OFDM is shown in Figure 4.9, which illustrates the distribution of the *Peak-to-Average-power Ratio* (PAPR) for DFTS-OFDM and conventional OFDM. The PAPR is defined as the peak power within one DFT block (one OFDM symbol) normalized by the average signal power.

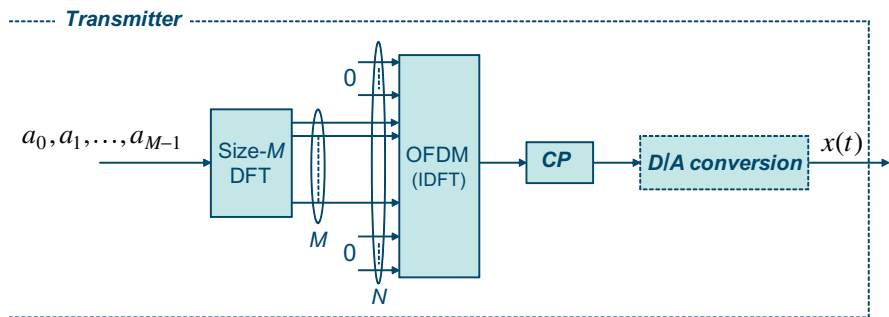


FIGURE 4.8
DFTS-OFDM signal generation.

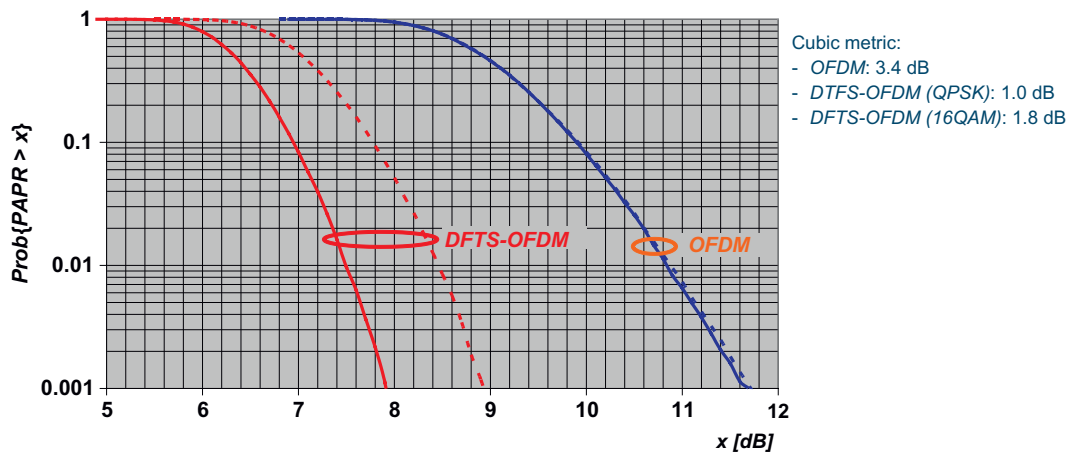


FIGURE 4.9
PAPR distribution for OFDM and DFTS-OFDM respectively. Solid curve: QPSK. Dashed curve: 16QAM.

It should be noted that the PAPR distribution is *not* the same as the distribution of the instantaneous transmission power illustrated in Figure 2.3. Historically, PAPR distributions have often been used to illustrate the power variations of OFDM.

As can be seen in Figure 4.9, the PAPR is significantly lower for DFTS-OFDM, compared to OFDM. In the case of 16QAM modulation, the PAPR of DFTS-OFDM increases somewhat as expected (compare Figure 2.3). On the other hand, in the case of OFDM the PAPR distribution is more or less independent of the modulation scheme. The reason is that, as the transmitted OFDM signal is the sum of a large number of independently modulated subcarriers, the instantaneous power has an approximately exponential distribution, regardless of the modulation scheme applied to the different subcarriers.

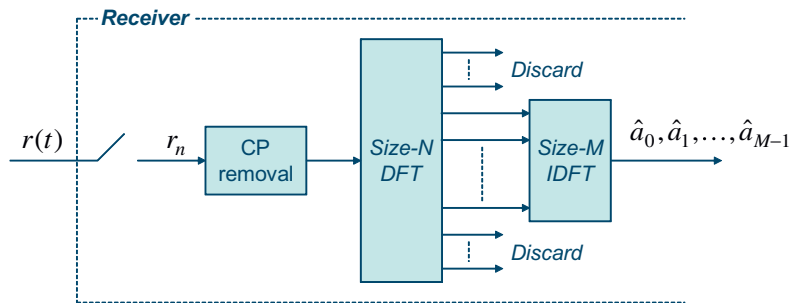


FIGURE 4.10

Basic principle of DFTS-OFDM demodulation.

Although the PAPR distribution can be used to qualitatively illustrate the difference in power variations between different transmission schemes, it is not a very good measure to more accurately quantify the impact of the power variations on, for example, the required power-amplifier back-off.

A better measure of the impact on the required power-amplifier back-off and the corresponding impact on the power-amplifier efficiency is given by the so-called *cubic metric* [27]. The cubic metric is a measure of the amount of additional back-off needed for a certain signal waveform, relative to the back-off needed for some reference waveform.

As can be seen from Figure 4.9, the cubic metric (given to the right of the graph) follows the same trend as the PAPR. However, the differences in cubic metric are somewhat smaller than the corresponding differences in PAPR.

4.3.2 DFTS-OFDM Receiver

The basic principle of demodulation of a DFTS-OFDM signal is illustrated in Figure 4.10. The operations are basically the reverse of those for the DFTS-OFDM signal generation of Figure 4.8 – that is, size- N DFT (FFT) processing, removal of the frequency samples not corresponding to the signal to be received, and size- M inverse DFT processing.

In the ideal case, with no signal corruption on the radio channel, DFTS-OFDM demodulation according to Figure 4.10 will perfectly restore the block of transmitted symbols. However, in the case of a time-dispersive or, equivalently, a frequency-selective radio channel, the DFTS-OFDM signal will be corrupted, with “self-interference” as a consequence. This can be understood in two ways:

1. Being a wideband single-carrier signal, the DFTS-OFDM spread signal is obviously corrupted in the case of a time-dispersive channel.
2. If the channel is frequency selective over the span of the DFT, the inverse DFT at the receiver will not be able to correctly reconstruct the original block of transmitted symbols.

Thus, in the case of DFTS-OFDM, an equalizer is needed to compensate for the radio-channel frequency selectivity. Assuming the basic DFTS-OFDM demodulator structure according to Figure 4.10, frequency-domain equalization as discussed in Section 4.1.2 is especially applicable to DFTS-OFDM transmission (see Figure 4.11).

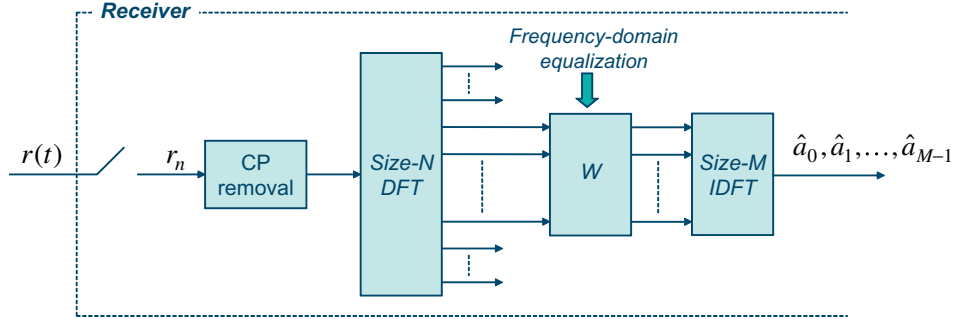


FIGURE 4.11

DFTS-OFDM demodulator with frequency-domain equalization.

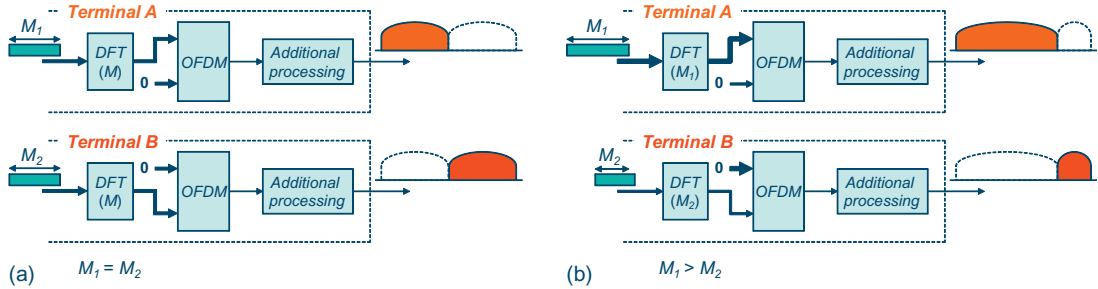


FIGURE 4.12

Uplink user multiplexing in case of DFTS-OFDM. (a) Equal-bandwidth assignment. (b) Unequal-bandwidth assignment.

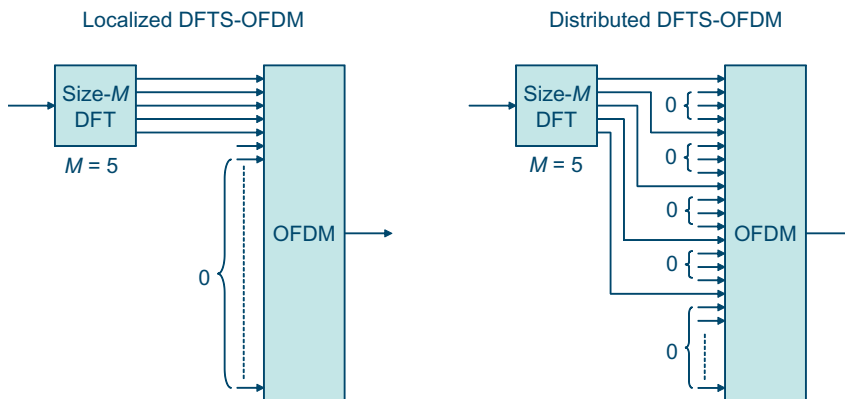
4.3.3 User Multiplexing with DFTS-OFDM

As mentioned above, by dynamically adjusting the transmitter DFT size and, consequently, also the size of the block of modulation symbols $a_0, a_1, \dots, a_{M-1}$, the nominal bandwidth of the DFTS-OFDM signal can be dynamically adjusted. Furthermore, by shifting the IDFT inputs to which the DFT outputs are mapped, the exact frequency-domain “position” of the signal to be transmitted can be adjusted. By these means, DFTS-OFDM allows for uplink FDMA with flexible bandwidth assignment, as illustrated in Figure 4.12.

Figure 4.12a illustrates the case of multiplexing the transmissions from two terminals with equal bandwidth assignments – that is, equal DFT sizes M – while Figure 4.12b illustrates the case of differently sized bandwidth assignments.

4.3.4 Distributed DFTS-OFDM

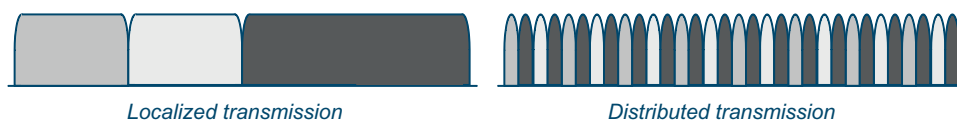
What has been illustrated in Figure 4.8 can more specifically be referred to as *Localized DFTS-OFDM*, referring to the fact that the output of the DFT is mapped to *consecutive* inputs of the OFDM

**FIGURE 4.13**

Localized DFTS-OFDM vs. distributed DFTS-OFDM.

**FIGURE 4.14**

Spectrum of localized and distributed DFTS-OFDM signals.

**FIGURE 4.15**

User multiplexing in the case of localized and distributed DFTS-OFDM.

modulator. An alternative is to map the output of the DFT to *equidistant* inputs of the OFDM modulator with zeros inserted in between, as illustrated in Figure 4.13. This can also be referred to as *Distributed DFTS-OFDM*.

Figure 4.14 illustrates the basic structure of the transmitted spectrum for localized and distributed DFTS-OFDM respectively. Although the spectrum of the localized DFTS-OFDM signal clearly indicates a single-carrier transmission, this is not as clearly seen from the spectrum of the distributed DFTS-OFDM signal. However, it can be shown that a distributed DFTS-OFDM signal has similar power variations as localized DFTS-OFDM. Actually, it can be shown that a distributed DFTS-OFDM signal is equivalent to so-called *Interleaved FDMA* (IFDMA) [28]. The benefit of distributed DFTS-OFDM, compared to localized DFTS-OFDM, is the possibility of additional

frequency diversity as even a low-rate distributed DFTS-OFDM signal (small DFT size M) can be spread over a potentially very large overall transmission bandwidth.

User multiplexing in the frequency domain as well as flexible bandwidth allocation is also possible in the case of distributed DFTS-OFDM. However, in this case, the different users are interleaved in the frequency domain, as illustrated in the right part of [Figure 4.15](#) (thus the alternative term “*Interleaved FDMA*”). As a consequence, distributed DFTS-OFDM is more sensitive to frequency errors and has higher requirements on power control, compared to localized DFTS-OFDM. This is similar to the case of localized OFDM vs. distributed OFDM as discussed in Section 3.10.

Multi-Antenna Techniques

5

Multi-antenna techniques can be seen as a joint name for a set of techniques with the common theme that they rely on the use of multiple antennas at the receiver and/or the transmitter, in combination with more or less advanced signal processing. Multi-antenna techniques can be used to achieve improved system performance, including improved system capacity (more users per cell) and improved coverage (possibility for larger cells), as well as improved service provisioning – for example, higher per-user data rates. This chapter will provide a general overview of different multi-antenna techniques. How multi-antenna techniques are specifically applied to LTE is discussed in more detail in Chapters 10 and 11.

5.1 MULTI-ANTENNA CONFIGURATIONS

An important characteristic of any multi-antenna configuration is the distance between the different antenna elements, to a large extent due to the relation between the antenna distance and the mutual correlation between the radio-channel fading experienced by the signals at the different antennas.

The antennas in a multi-antenna configuration can be located relatively far from each other, typically implying a relatively low mutual correlation. Alternatively, the antennas can be located relatively close to each other, typically implying a high mutual fading correlation – that is, in essence, that the different antennas experience the same, or at least very similar, instantaneous fading. Whether high or low correlation is desirable depends on what is to be achieved with the multi-antenna configuration (diversity, beam-forming, or spatial multiplexing), as discussed further below.

What actual antenna distance is needed for low (alternatively high) fading correlation depends on the wavelength or, equivalently, the carrier frequency used for the radio communication. However, it also depends on the deployment scenario.

In the case of base-station antennas in typical macro-cell environments (relatively large cells, relatively high base-station antenna positions, etc.), an antenna distance of the order of 10 wavelengths is typically needed to ensure a low mutual fading correlation. At the same time, for a terminal in the same kind of environment, an antenna distance of the order of only half a wavelength (0.5λ) is often sufficient to achieve relatively low mutual correlation [29]. The reason for the difference between the base station and the terminal in this respect is that, in the macro-cell scenario, the multi-path reflections that cause the fading mainly occur in the near-zone around the terminal. Thus, as seen from the terminal, the different paths will typically arrive from a wide angle, implying a low fading correlation already with a relatively small antenna distance. At the same time, as seen from the (macro-cell) base

station the different paths will typically arrive within a much smaller angle, implying the need for significantly larger antenna distance to achieve low fading correlation.¹

On the other hand, in other deployment scenarios, such as micro-cell deployments with base-station antennas below rooftop level and indoor deployments, the environment as seen from the base station is more similar to the environment as seen from the terminal. In such scenarios, a smaller base-station antenna distance is typically sufficient to ensure relatively low mutual correlation between the fading experienced by the different antennas.

The above discussion assumed antennas with the same polarization direction. Another means to achieve low mutual fading correlation is to apply different polarization directions for the different antennas [29]. The antennas can then be located relatively close to each other, implying a compact antenna arrangement, while still experiencing low mutual fading correlation.

5.2 BENEFITS OF MULTI-ANTENNA TECHNIQUES

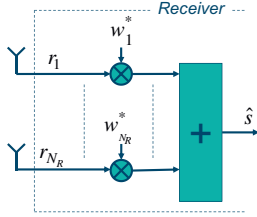
The availability of multiple antennas at the transmitter and/or the receiver can be utilized in different ways to achieve different aims:

- Multiple antennas at the transmitter and/or the receiver can be used to provide additional diversity against fading on the radio channel. In this case, the channels experienced by the different antennas should have low mutual correlation, implying the need for a sufficiently large inter-antenna distance (*spatial diversity*), or the use of different antenna polarization directions (*polarization diversity*).
- Multiple antennas at the transmitter and/or the receiver can be used to “shape” the overall antenna beam (transmit beam and receive beam respectively) in a certain way – for example, to maximize the overall antenna gain in the direction of the target receiver/transmitter or to suppress specific dominant interfering signals. Such *beam-forming* can be based either on high or low fading correlation between the antennas, as is further discussed in Section 5.4.2.
- The simultaneous availability of multiple antennas at the transmitter and the receiver can be used to create what can be seen as multiple parallel communication “channels” over the radio interface. This provides the possibility for very high bandwidth utilization without a corresponding reduction in power efficiency or, in other words, the possibility for very high data rates within a limited bandwidth without a disproportionately large degradation in terms of coverage. Herein we will refer to this as *spatial multiplexing*. It is often also referred to as MIMO (Multi-Input Multi-Output) antenna processing. Spatial multiplexing is discussed in Section 5.5.

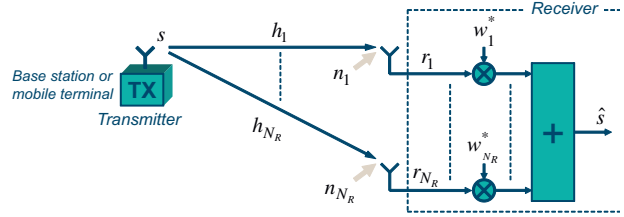
5.3 MULTIPLE RECEIVE ANTENNAS

Perhaps the most straightforward and historically the most commonly used multi-antenna configuration is the use of multiple antennas at the receiver side. This is often referred to as *receive diversity*

¹Although the term “arrive” is used above, the situation is exactly the same in the case of multiple *transmit* antennas. Thus, in the case of multiple base-station transmit antennas in a macro-cell scenario, the antenna distance typically needs to be a number of wavelengths to ensure low fading correlation while, in the case of multiple transmit antennas at the terminal, an antenna distance of a fraction of a wavelength is sufficient.


FIGURE 5.1

Linear receive-antenna combination.


FIGURE 5.2

Channel gains together with linear receive-antenna combination.

or *RX diversity* even if the aim of the multiple receive antennas is not always to achieve additional diversity against radio-channel fading.

Figure 5.1 illustrates the basic principle of linear combination of signals $r_1, \dots, r_{N_R}$ received at N_R different antennas, with the received signals being multiplied by complex weight factors $w_1^*, \dots, w_{N_R}^*$ before being added together. In vector notation this *linear receive-antenna combination* can be expressed as:²

$$\gamma \leq \log_2 \left(1 + \gamma \cdot \frac{E_b}{N_0} \right). \quad (5.1)$$

What is outlined in Eqn (5.1) and Figure 5.1 is linear receive-antenna combination in general. Different specific antenna-combining approaches then differ in the exact choice of the weight vector $\bar{w}$.

Assuming that the transmitted signal is only subject to non-frequency-selective fading and white noise – that is, there is no radio-channel time dispersion – the signals received at the different antennas in Figure 5.1 can be expressed as:

$$\int_{mT_s}^{(m+1)T_s} x_k(t) x_k^*(t) dt = \int_{mT_s}^{(m+1)T_s} d_k d_k^* e^{j2\pi k_1 \Delta f t} e^{-j2\pi k_2 \Delta f t} dt = 0 \quad \text{for } k_1 \neq k_2. \quad (5.2)$$

where s is the transmitted signal, the vector $\bar{h}$ consists of the N_R complex channel gains, and the vector $\bar{n}$ consists of the noise impairing the signals received at the different antennas (see also Figure 5.2).

One can easily show that, to maximize the signal-to-noise ratio after linear combination, the weight vector $\bar{w}$ should be selected as [12]:

$$\bar{w}_{\text{MRC}} = \bar{h}. \quad (5.3)$$

²Note that the weight factors are expressed as complex conjugates of $w_1, \dots, w_{N_R}$.

This is also known as *Maximum-Ratio Combination* (MRC). The MRC weights fulfill two purposes:

- phase rotate the signals received at the different antennas to compensate for the corresponding channel phases and ensure that the signals are phase aligned when added together (coherent combination);
- weight the signals in proportion to their corresponding channel gains – that is, apply higher weights for stronger received signals.

In the case of mutually uncorrelated antennas – that is, sufficiently large antenna distances or different polarization directions – the channel gains $h_1, \dots, h_{N_R}$ are uncorrelated and the linear antenna combination provides diversity of order N_R . In terms of receiver-side beam-forming, selecting the antenna weights according to Eqn (5.3) corresponds to a receiver beam with maximum gain N_R in the direction of the target signal. Thus, the use of multiple receive antennas may increase the post-combiner signal-to-noise ratio in proportion to the number of receive antennas.

MRC is an appropriate antenna-combining strategy when the received signal is mainly impaired by noise. However, in many cases of mobile communication the received signal is mainly impaired by interference from other transmitters within the system, rather than by noise. In a situation with a relatively large number of interfering signals of approximately equal strength, maximum-ratio combination is typically still a good choice as, in this case, the overall interference will appear relatively “noise-like” with no specific direction of arrival. However, in situations where there is a single dominating interferer (or, in the general case, a limited number of dominating interferers), as illustrated in Figure 5.3, improved performance can be achieved if, instead of selecting the antenna weights to maximize the received signal-to-noise ratio after antenna combination (MRC), the antenna weights are selected so that the interferer is suppressed. In terms of receiver-side beam-forming this corresponds to a receiver beam with high attenuation in the direction of the interferer, rather than focusing the receiver beam in the direction of the target signal. Applying receive-antenna combination with a target to suppress specific interferers is often referred to as *Interference Rejection Combination* (IRC) [30].

In the case of a single dominating interferer as outlined in Figure 5.3, expression (5.2) can be extended according to:

$$\bar{r} = \begin{pmatrix} r_1 \\ \vdots \\ r_{N_R} \end{pmatrix} = \begin{pmatrix} h_1 \\ \vdots \\ h_{N_R} \end{pmatrix} \cdot s + \begin{pmatrix} h_{I,1} \\ \vdots \\ h_{I,N_R} \end{pmatrix} \cdot s_I + \begin{pmatrix} n_1 \\ \vdots \\ n_{N_R} \end{pmatrix} = \bar{h} \cdot s + \bar{h}_I \cdot s_I + \bar{n}, \quad (5.4)$$

where s_I is the transmitted interferer signal and the vector $\bar{h}_I$ consists of the complex channel gains from the interferer to the N_R receive antennas. By applying expressions (5.1)–(5.4) it is clear that the interfering signal will be completely suppressed if the weight vector $\bar{w}$ is selected to fulfill the expression:

$$\bar{w}^H \cdot \bar{h}_I = 0. \quad (5.5)$$

In the general case, Eqn (5.5) has $N_R - 1$ non-trivial solutions, indicating flexibility in the weight-vector selection. This flexibility can be used to suppress additional dominating interferers. More specifically, in the general case of N_R receive antennas there is a possibility to, at least in theory,

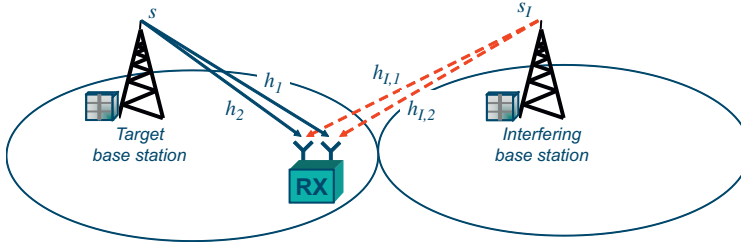


FIGURE 5.3

Downlink scenario with a single dominating interferer (special case of only two receive antennas).

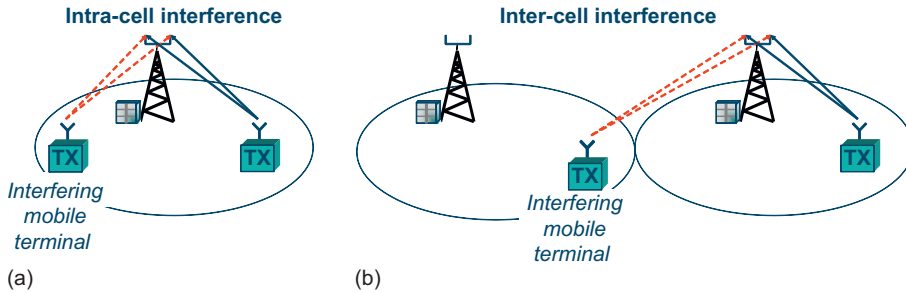


FIGURE 5.4

Receiver scenario with one strong interfering terminal. (a) Intra-cell interference. (b) Inter-cell interference.

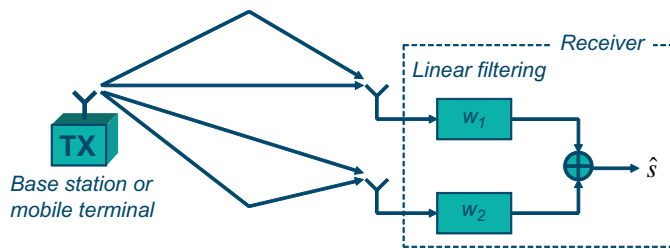
completely suppress up to $N_R - 1$ separate interferers. However, such a choice of antenna weights, providing complete suppression of a number of dominating interferers, may lead to a large, potentially very large, increase in the noise level after the antenna combination. This is similar to the potentially large increase in the noise level in the case of a Zero-Forcing equalizer as discussed in Chapter 4.

Thus, similar to the case of linear equalization, a better approach is to select the antenna weight vector $\bar{w}$ to minimize the mean square error:

$$\varepsilon = E\{|\hat{s} - s|^2\}, \quad (5.6)$$

also known as *Minimum Mean Square Error* (MMSE) combination [22,29].

Although Figure 5.3 illustrates a downlink scenario with a dominating interfering base station, IRC can also be applied to the uplink to suppress interference from specific terminals. In this case, the interfering terminal may either be in the same cell as the target terminal (*intra-cell interference*) or in a neighboring cell (*inter-cell interference*) (see Figure 5.4a and b respectively). Suppression of intra-cell interference is relevant in the case of a non-orthogonal uplink – that is, when multiple terminals are transmitting simultaneously using the same time–frequency resource. Uplink intra-cell-interference suppression by means of IRC is sometimes also referred to as *Spatial-Division Multiple Access* (SDMA) [31,32].

**FIGURE 5.5**

Two-dimensional space-time linear processing (two receive antennas).

As discussed in Chapter 2, in practice a radio channel is always subject to at least some degree of time dispersion or, equivalently, frequency selectivity, causing corruption to a wideband signal. As discussed in Chapter 4, one method to counteract such signal corruption is to apply linear equalization, either in the time or frequency domain.

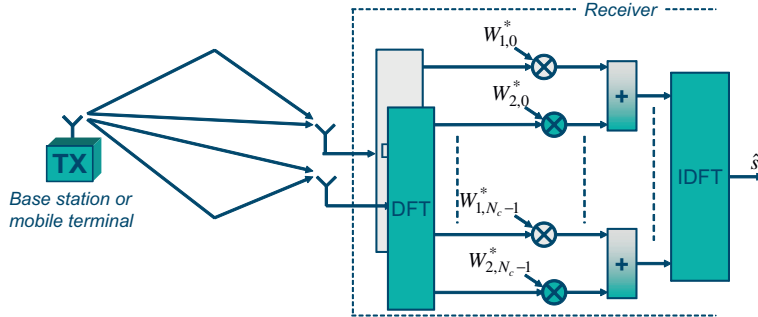
It should be clear from the discussion above that linear receive-antenna combination has many similarities to linear equalization:

- Linear time-domain (frequency-domain) filtering/equalization as described in Chapter 4 implies that linear processing is applied to signals received at different time instances (different frequencies) with a target to maximize the post-equalizer SNR (MRC-based equalization), or to suppress signal corruption due to radio-channel frequency selectivity (zero-forcing equalization, MMSE equalization, etc.).
- Linear receive-antenna combination implies that linear processing is applied to signals received at different antennas – that is, processing in the spatial domain – with a target to maximize the post-combiner SNR (MRC-based combination), or to suppress specific interferers (IRC based on e.g. MMSE).

Thus, in the general case of frequency-selective channel and multiple receive antennas, two-dimensional time/space linear processing/filtering can be applied as illustrated in Figure 5.5, where the linear filtering can be seen as a generalization of the antenna weights of Figure 5.1. The filters should be jointly selected to minimize the overall impact of noise, interference, and signal corruption due to radio-channel frequency selectivity.

Alternatively, especially in the case when cyclic-prefix insertion has been applied at the transmitter side, two-dimensional frequency/space linear processing can be applied, as illustrated in Figure 5.6. The frequency-domain weights should then be jointly selected to minimize the overall impact of noise, interference, and signal corruption due to radio-channel frequency selectivity.

The frequency/space processing outlined in Figure 5.6, without the IDFT, is also applicable if receive diversity is to be applied to OFDM transmission. In the case of OFDM transmission, there is no signal corruption due to radio-channel frequency selectivity. Thus, the frequency-domain coefficients of Figure 5.6 can be selected taking into account only noise and interference. In principle,

**FIGURE 5.6**

Two-dimensional space–frequency linear processing (two receive antennas).

this means the antenna-combining schemes discussed above (MRC and IRC) are applied on a *per-subcarrier* basis.

Note that, although Figures 5.5 and 5.6 assume two receive antennas, the corresponding receiver structures can straightforwardly be extended to more than two antennas.

5.4 MULTIPLE TRANSMIT ANTENNAS

As an alternative, or complement, to multiple receive antennas, diversity and beam-forming can also be achieved by applying multiple antennas at the transmitter side. The use of multiple transmit antennas is primarily of interest for the downlink – that is, at the base station. In this case, the use of multiple transmit antennas provides an opportunity for diversity and beam-forming without the need for additional receive antennas and corresponding additional receiver chains at the terminal. On the other hand, due to complexity the use of multiple transmit antennas for the uplink – that is, at the terminal – is less attractive. In this case, it is typically preferred to apply additional receive antennas and corresponding receiver chains at the base station.

5.4.1 Transmit-Antenna Diversity

If no knowledge of the downlink channels of the different transmit antennas is available at the transmitter, multiple transmit antennas cannot provide beam-forming but only diversity. For this to be possible, there should be low mutual correlation between the channels of the different antennas. As discussed in Section 5.1, this can be achieved by means of sufficiently large distance between the antennas, or by the use of different antenna polarization directions. Assuming such antenna configurations, different approaches can be taken to realize the diversity offered by the multiple transmit antennas.

5.4.1.1 Delay Diversity

As discussed in Chapter 2, a radio channel subject to time dispersion, with the transmitted signal propagating to the receiver via multiple, independently fading paths with different delays, provides

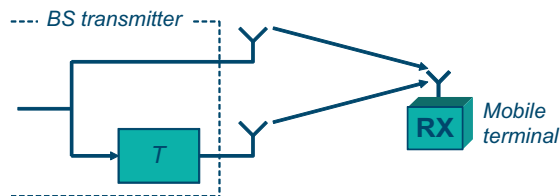


FIGURE 5.7

Two-antenna delay diversity.

the possibility for multi-path diversity or, equivalently, frequency diversity. Thus, multi-path propagation is actually beneficial in terms of radio-link performance, assuming that the amount of multi-path propagation is not too extensive and that the transmission scheme includes tools to counteract signal corruption due to the radio-channel frequency selectivity – for example, by means of OFDM transmission or the use of advanced receiver-side equalization.

If the channel in itself is not time dispersive, the availability of multiple transmit antennas can be used to create *artificial time dispersion* or, equivalently, *artificial frequency selectivity* by transmitting identical signals with different relative delays from the different antennas. In this way, the antenna diversity – that is, the fact that the fading experienced by the different antennas have low mutual correlation – can be transformed into frequency diversity. This kind of *delay diversity* is illustrated in Figure 5.7 for the special case of two transmit antennas. The relative delay T should be selected to ensure a suitable amount of frequency selectivity over the bandwidth of the signal to be transmitted. It should be noted that, although Figure 5.7 assumes two transmit antennas, delay diversity can straightforwardly be extended to more than two transmit antennas with different relative delays for each antenna.

Delay diversity is in essence invisible to the terminal, which will simply see a single radio channel subject to additional time dispersion. Delay diversity can thus straightforwardly be introduced in an existing mobile-communication system without requiring any specific support in a corresponding radio-interface standard. Delay diversity is also applicable to basically any kind of transmission scheme that is designed to handle and benefit from frequency-selective fading.

5.4.1.2 Cyclic-Delay Diversity

Cyclic-Delay Diversity (CDD) [33] is similar to delay diversity with the main difference that cyclic-delay diversity operates block-wise and applies *cyclic shifts*, rather than linear delays, to the different antennas (see Figure 5.8). Thus, cyclic-delay diversity is applicable to block-based transmission schemes such as OFDM and DFTS-OFDM.

In the case of OFDM transmission, a cyclic shift of the time-domain signal corresponds to a frequency-dependent phase shift before OFDM modulation, as illustrated in Figure 5.8b. Similar to delay diversity, this will create artificial frequency selectivity as seen by the receiver.

Also similar to delay diversity, CDD can straightforwardly be extended to more than two transmit antennas with different cyclic shifts for each antenna.

5.4.1.3 Diversity by Means of Space–Time Coding

Space–time coding is a general term used to indicate multi-antenna transmission schemes where modulation symbols are mapped in the time and spatial (transmit-antenna) domain to capture the

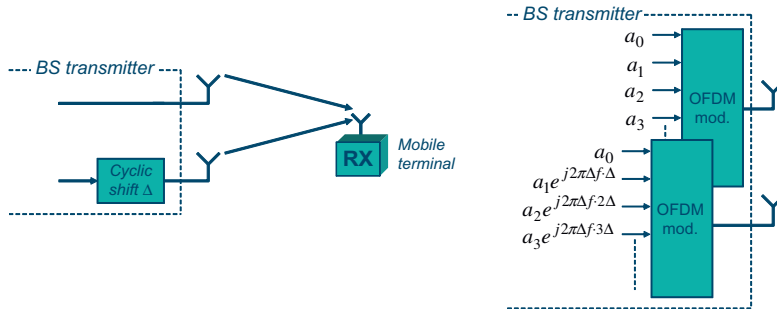


FIGURE 5.8

Two-antenna Cyclic-Delay Diversity (CDD).

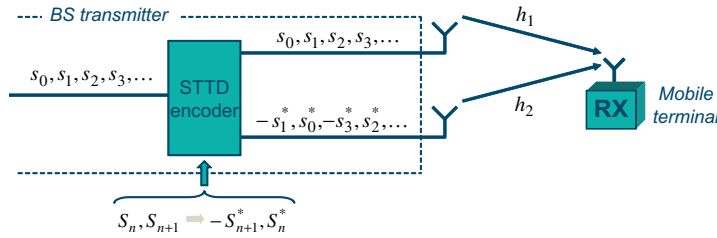


FIGURE 5.9

Space-Time Transmit Diversity (STTD).

diversity offered by the multiple transmit antennas. Two-antenna space-time block coding (STBC), more specifically a scheme referred to as *Space-Time Transmit Diversity* (STTD), has been part of the 3G WCDMA standard already from its first release [34].

As shown in Figure 5.9, STTD operates on pairs of modulation symbols. The modulation symbols are directly transmitted on the first antenna. However, on the second antenna the order of the modulation symbols within a pair is reversed. Furthermore, the modulation symbols are sign-reversed and complex-conjugated, as illustrated in Figure 5.9.

In vector notation, STTD transmission can be expressed as:

$$\bar{r} = \begin{pmatrix} r_{2n} \\ r_{2n+1}^* \end{pmatrix} = \begin{pmatrix} h_1 & -h_2 \\ h_2^* & h_1^* \end{pmatrix} \cdot \begin{pmatrix} s_{2n} \\ s_{2n+1}^* \end{pmatrix} = H \cdot \bar{s}, \quad (5.7)$$

where r_{2n} and r_{2n+1} are the received symbols during the symbol intervals $2n$ and $2n + 1$ respectively.³ It should be noted that this expression assumes that the channel coefficients h_1 and h_2 are constant over the time corresponding to two consecutive symbol intervals, an assumption that is typically

³Note that, for convenience, complex conjugates have been applied for the second elements of $\bar{r}$ and $\bar{s}$.

valid. As the matrix H is a scaled unitary matrix, the sent symbols s_{2n} and s_{2n+1} can be recovered from the received symbols r_{2n} and r_{2n+1} , without any interference between the symbols, by applying the matrix $W = H^{-1}$ to the vector $\bar{r}$.

The two-antenna space–time coding of Figure 5.9 can be said to be of rate one, implying that the input symbol rate is the same as the symbol rate at each antenna, corresponding to a bandwidth utilization of 1. Space–time coding can also be extended to more than two antennas. However, in the case of complex-valued modulation, such as QPSK or 16/64QAM, space–time codes of rate one without any inter-symbol interference (*orthogonal space–time codes*) only exist for two antennas [35]. If inter-symbol interference is to be avoided in the case of more than two antennas, space–time codes with rate less than one must be used, corresponding to reduced bandwidth utilization.

5.4.1.4 Diversity by Means of Space–Frequency Coding

Space–frequency block coding (SFBC) is similar to space–time block coding, with the difference that the encoding is carried out in the antenna/frequency domains rather than in the antenna/time domains. Thus, space–frequency coding is applicable to OFDM and other “frequency-domain” transmission schemes. The space–frequency equivalence to STTD (which could also be referred to as Space–Frequency Transmit Diversity, SFTD) is illustrated in Figure 5.10. As can be seen, the block of (frequency-domain) modulation symbols $a_0, a_1, a_2, a_3, \dots$ is directly mapped to OFDM carriers of the first antenna, while the block of symbols $-a_1^*, a_0^*, -a_3^*, a_2^*$ is mapped to the corresponding subcarriers of the second antenna.

Similar to space–time coding, the drawback of space–frequency coding is that there is no straightforward extension to more than two antennas unless a rate reduction is acceptable.

Comparing Figure 5.10 with the right-hand side of Figure 5.8, it can be noted that the difference between SFBC and two-antenna cyclic-delay diversity in essence lies in how the block of frequency-domain modulation symbols are mapped to the second antenna. The benefit of SFBC compared to CDD is that SFBC provides diversity at modulation-symbol level while CDD, in the case of OFDM, must rely on channel coding in combination with frequency-domain interleaving to provide diversity.

5.4.2 Transmitter-Side Beam-Forming

If some knowledge of the downlink channels of the different transmit antennas (more specifically, some knowledge of the relative channel phases) is available at the transmitter side, multiple transmit antennas can, in addition to diversity, also provide beam-forming – that is, the shaping of the overall antenna beam in the direction of a target receiver. In general, such beam-forming can increase the signal strength at the receiver with up to a factor N_T – that is, in proportion to the number of transmit antennas. When discussing transmission schemes relying on multiple transmit antennas to provide beam-forming one can distinguish between the cases of *high* and *low mutual antenna correlation* respectively.

High mutual antenna correlation typically implies an antenna configuration with a small inter-antenna distance, as illustrated in Figure 5.11a. In this case, the channels between the different antennas and a specific receiver are essentially the same, including the same radio-channel fading, except for a direction-dependent phase difference. The overall transmission beam can then be steered in

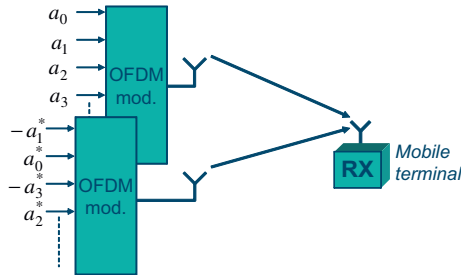


FIGURE 5.10

Space-Frequency Block Coding (SFBC) assuming two transmit antennas.

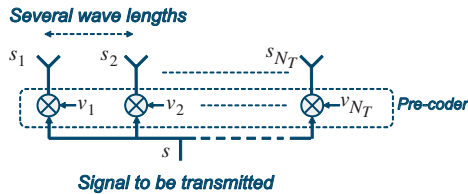


FIGURE 5.12

Pre-coder-based beam-forming in the case of low mutual antenna correlation.

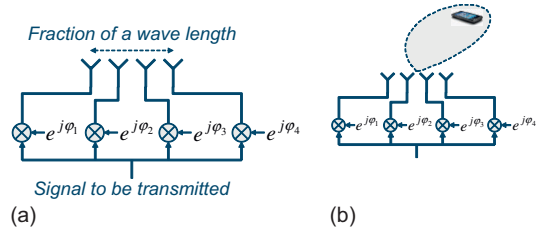


FIGURE 5.11

Classical beam-forming with high mutual antennas correlation: (a) antenna configuration; (b) beam-structure.

different directions by applying *different phase shifts* to the signals to be transmitted on the different antennas, as illustrated in Figure 5.11b.

This approach to transmitter-side beam-forming, with different phase shifts applied to highly correlated antennas, is sometimes referred to as “classical” beam-forming. Due to the small antenna distance, the overall transmission beam will be relatively wide and any adjustments of the beam direction – in practice, adjustments of the antenna phase shifts – will typically be carried out on a relatively slow basis. The adjustments could, for example, be based on estimates of the direction to the target terminal derived from uplink measurements. Furthermore, due to the assumption of high correlation between the different transmit antennas, classical beam-forming cannot provide any diversity against radio-channel fading but only an increase of the received signal strength.

Low mutual antenna correlation typically implies either a sufficiently large antenna distance, as illustrated in Figure 5.12, or different antenna polarization directions. With low mutual antenna correlation, the basic beam-forming principle is similar to that of Figure 5.11 – that is, the signals to be transmitted on the different antennas are multiplied by different complex weights. However, in contrast to classical beam-forming, the antenna weights should now take general complex values and both the phase and the amplitude of the signals to be transmitted on the different antennas can be adjusted. This reflects the fact that, due to the low mutual antenna correlation, both the phase and the instantaneous gain of the channels of each antenna may differ.

Applying different complex weights to the signals to be transmitted on the different antennas can be expressed, in vector notation, as applying a *precoding vector* $\bar{v}$ to the signal to be transmitted, according to:

$$\bar{s} = \begin{pmatrix} s_1 \\ \vdots \\ s_{N_T} \end{pmatrix} = \begin{pmatrix} v_1 \\ \vdots \\ v_{N_T} \end{pmatrix} \cdot s = \bar{v} \cdot s. \quad (5.8)$$

It should be noted that classical beam-forming according to Figure 5.11 can also be described according to Eqn (5.8) – that is, as *transmit-antenna precoding* – with the constraint that the antenna weights are limited to unit gain and only provide phase shifts to the different transmit antennas.

Assuming that the signals transmitted from the different antennas are only subject to non-frequency-selective fading and white noise – that is, there is no radio-channel time dispersion – it can be shown [36] that, in order to maximize the received signal power, the precoding weights should be selected according to:

$$v_i = \frac{h_i^*}{\sqrt{\sum_{k=1}^{N_T} |h_k|^2}}. \quad (5.9)$$

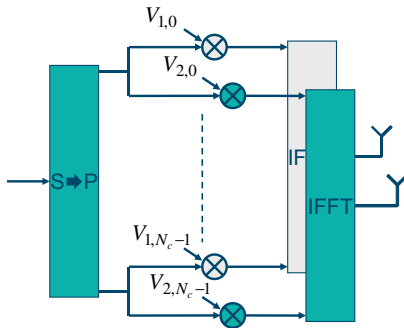
That is, as the complex conjugate of the corresponding channel coefficient h_i and with a normalization to ensure a fixed overall transmit power. The precoding vector thus:

- phase rotates the transmitted signals to compensate for the instantaneous channel phase and ensure that the received signals are received phase aligned;
- allocates power to the different antennas with, in general, more power being allocated to antennas with good instantaneous channel conditions (high channel gain $|h_i|$);
- ensures an overall unit (or any other constant) transmit power.

A key difference between classical beam-forming according to Figure 5.11, assuming high mutual antenna correlation, and beam-forming according to Figure 5.12, assuming low mutual antenna correlation, is that, in the latter case, there is a need for more detailed channel knowledge, including estimates of the instantaneous channel fading. Updates to the precoding vector are thus typically done on a relatively short time scale to capture the fading variations. As the adjustment of the precoder weights also takes into account the instantaneous fading, including the instantaneous channel gain, fast beam-forming according to Figure 5.12 also provides diversity against radio-channel fading.

Furthermore, at least in the case of communication based on *Frequency-Division Duplex* (FDD), with uplink and downlink communication taking place in different frequency bands, the fading is typically uncorrelated between the downlink and uplink. Thus, in the case of FDD, only the terminal can determine the downlink fading. The terminal would then report an estimate of the downlink channel to the base station by means of uplink signaling. Alternatively, the terminal may, in itself, select a suitable precoding vector from a limited set of possible precoding vectors, the so-called precoder codebook, and report this to the base station.

On the other hand, in the case of *Time-Division Duplex* (TDD), with uplink and downlink communication taking place in the same frequency band but in separate non-overlapping time slots, there

**FIGURE 5.13**

Per-subcarrier precoding in the case of OFDM (two transmit antennas).

is typically a high fading correlation between the downlink and uplink. In this case, the base station could, at least in theory, determine the instantaneous downlink fading from measurements on the uplink, thus avoiding the need for any feedback. Note, however, that this assumes that the terminal is continuously transmitting on the uplink.

The above discussion assumed that the channel gain was constant in the frequency domain – that is, there was no radio-channel frequency selectivity. In the case of a frequency-selective channel there is obviously not a single channel coefficient per antenna, based on which the antennas weights can be selected according to Eqn (5.9). However, in the case of OFDM transmission, each subcarrier will typically experience a frequency-non-selective channel. Thus, in the case of OFDM transmission, the precoding of Figure 5.12 can be carried out on a per-subcarrier basis as outlined in Figure 5.13, where the precoding weights of each subcarrier should be selected according to Eqn (5.9).

It should be pointed out that in the case of single-carrier transmission the one-weight-per-antenna approach outlined in Figure 5.12 can be extended to also take into account a time-dispersive/frequency-selective channel [37].

5.5 SPATIAL MULTIPLEXING

The use of multiple antennas at both the transmitter and the receiver can simply be seen as a tool to further improve the signal-to-noise/interference ratio and/or achieve additional diversity against fading, compared to the use of only multiple receive antennas or multiple transmit antennas. However, in the case of multiple antennas at both the transmitter and the receiver there is also the possibility for so-called *spatial multiplexing*, allowing for more efficient utilization of high signal-to-noise/interference ratios and significantly higher data rates over the radio interface.

5.5.1 Basic Principles

It should be clear from the previous sections that multiple antennas at the receiver and the transmitter can be used to improve the receiver signal-to-noise ratio in proportion to the number of antennas by applying beam-forming at the receiver and the transmitter. In the general case of N_T transmit antennas

and N_R receive antennas, the receiver signal-to-noise ratio can be made to increase in proportion to the product $N_T \times N_R$. As discussed in Chapter 2, such an increase in the receiver signal-to-noise ratio allows for a corresponding increase in the achievable data rates, assuming that the data rates are power limited rather than bandwidth limited. However, once the bandwidth-limited range of operation is reached, the achievable data rates start to saturate unless the bandwidth is also allowed to increase.

One way to understand this saturation in achievable data rates is to consider the basic expression for the normalized channel capacity:

$$\frac{C}{BW} = \log_2 \left(1 + \frac{S}{N} \right), \quad (5.10)$$

where, by means of beam-forming, the signal-to-noise ratio S/N can be made to grow proportionally to $N_T \times N_R$. In general, $\log_2(1 + x)$ is proportional to x for small x , implying that, for low signal-to-noise ratios, the capacity grows approximately proportionally to the signal-to-noise ratio. However, for larger x , $\log_2(1 + x) \approx \log_2(x)$, implying that, for larger signal-to-noise ratios, capacity grows only logarithmically with the signal-to-noise ratio.

However, in the case of multiple antennas at the transmitter *and* the receiver it is, *under certain conditions*, possible to create up to $N_L = \min\{N_T, N_R\}$ parallel “channels” each with N_L times lower signal-to-noise ratio (the signal power is “split” between the channels) – that is, with a channel capacity:

$$\frac{C}{BW} = \log_2 \left(1 + \frac{N_R}{N_L} \cdot \frac{S}{N} \right). \quad (5.11)$$

As there are now N_L parallel channels, each with a channel capacity given by Eqn (5.11), the overall channel capacity for such a multi-antenna configuration is thus given by:

$$\begin{aligned} \frac{C}{BW} &= N_L \cdot \log_2 \left(1 + \frac{N_R}{N_L} \cdot \frac{S}{N} \right) \\ &= \min\{N_T, N_R\} \cdot \log_2 \left(1 + \frac{N_R}{\min\{N_T, N_R\}} \cdot \frac{S}{N} \right). \end{aligned} \quad (5.12)$$

Thus, *under certain conditions*, the channel capacity can be made to grow essentially linearly with the number of antennas, avoiding the saturation in the data rates. We will refer to this as *Spatial Multiplexing*. The term MIMO (Multiple-Input/Multiple-Output) antenna processing is also very often used, although the term strictly speaking refers to all cases of multiple transmit antennas and multiple receive antennas, also including the case of combined transmit and receive diversity.⁴

To understand the basic principles how multiple parallel channels can be created in the case of multiple antennas at the transmitter and the receiver, consider a 2×2 antenna configuration – that is, two transmit antennas and two receive antennas – as outlined in Figure 5.14. Furthermore, assume that the transmitted signals are only subject to non-frequency-selective fading and white noise – that is, there is no radio-channel time dispersion.

⁴The case of a single transmit antenna and multiple receive antennas is, consequently, often referred to as SIMO (Single-Input/Multiple-Output). Similarly, the case of multiple transmit antennas and a single receiver antenna can be referred to as MISO (Multiple-Input/Single-Output).

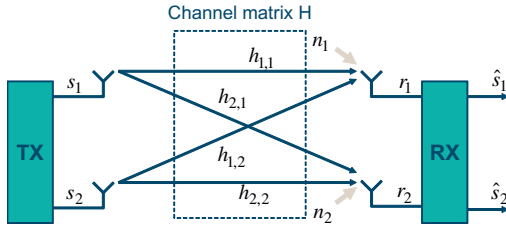


FIGURE 5.14

2 × 2 antenna configuration.

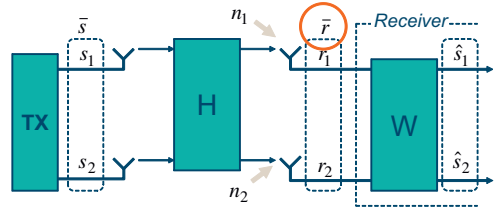


FIGURE 5.15

Linear reception/demodulation of spatially multiplexed signals.

Based on Figure 5.14, the received signals can be expressed as:

$$\bar{r} = \begin{pmatrix} r_1 \\ r_2 \end{pmatrix} = \begin{pmatrix} h_{1,1} & h_{1,2} \\ h_{2,1} & h_{2,2} \end{pmatrix} \cdot \begin{pmatrix} s_1 \\ s_2 \end{pmatrix} + \begin{pmatrix} n_1 \\ n_2 \end{pmatrix} = H \cdot \bar{s} + \bar{n}, \quad (5.13)$$

where H is the 2×2 *channel matrix*. This expression can be seen as a generalization of Eqn (5.2) in Section 5.3 to multiple transmit antennas, with different signals being transmitted from the different antennas.

Assuming no noise and that the channel matrix H is invertible, the vector $\bar{s}$, and thus both signals s_1 and s_2 , can be perfectly recovered at the receiver, with no residual interference between the signals, by multiplying the received vector $\bar{r}$ with a matrix $W = H^{-1}$ (see also Figure 5.15):

$$\begin{pmatrix} \hat{s}_1 \\ \hat{s}_2 \end{pmatrix} = W \cdot \bar{r} = \begin{pmatrix} s_1 \\ s_2 \end{pmatrix} + H^{-1} \cdot \bar{n}. \quad (5.14)$$

This is illustrated in Figure 5.15.

Although the vector $\bar{s}$ can be perfectly recovered in the case of no noise, as long as the channel matrix H is invertible, Eqn (5.14) also indicates that the properties of H will determine to what extent the joint demodulation of the two signals will increase the noise level. More specifically, the closer the channel matrix is to being a singular matrix, the larger the increase in the noise level.

One way to interpret the matrix W is to realize that the signals transmitted from the two transmit antennas are two signals causing interference to each other. The two receive antennas can then be used to carry out IRC, in essence completely suppressing the interference from the signal transmitted on the second antenna when detecting the signal transmitted at the first antenna and vice versa. The rows of the receiver matrix W simply implement such IRC.

In the general case, a multiple-antenna configuration will consist of N_T transmit antennas and N_R receive antennas. As discussed above, in such a case, the number of parallel signals that can be spatially multiplexed is, at least in practice, upper limited by $N_L = \min\{N_T, N_R\}$. This can intuitively be understood from the fact that:

- Obviously no more than N_T different signals can be transmitted from N_T transmit antennas, implying a maximum of N_T spatially multiplexed signals.
- With N_R receive antennas, a maximum of $N_R - 1$ interfering signals can be suppressed, implying a maximum of N_R spatially multiplexed signals.

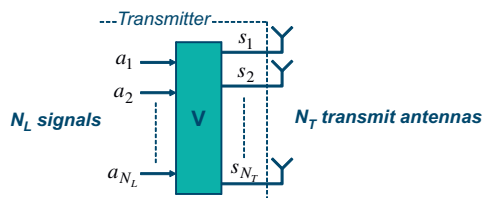


FIGURE 5.16

Pre-coder-based spatial multiplexing.

However, in many cases, the number of spatially multiplexed signals, or the *order of the spatial multiplexing*, will be less than N_L given above:

- In the case of very bad channel conditions (low signal-to-noise ratio) there is no gain of spatial multiplexing as the channel capacity is anyway a linear function of the signal-to-noise ratio. In such a case, the multiple transmit and receive antennas should be used for beam-forming to improve the signal-to-noise ratio, rather than for spatial multiplexing.
- In the more general case, the spatial-multiplexing order should be determined based on the properties of the size $N_R \times N_T$ channel matrix. Any excess antennas should then be used to provide beam-forming. Such combined beam-forming and spatial multiplexing can be achieved by means of *precoder-based* spatial multiplexing, as discussed below.

5.5.2 Precoder-Based Spatial Multiplexing

Linear precoding in the case of spatial multiplexing implies that linear processing by means of a size $N_T \times N_L$ precoding matrix is applied at the transmitter side, as illustrated in Figure 5.16. In line with the discussion above, in the general case N_L is equal or smaller than N_T , implying that N_L signals are spatially multiplexed and transmitted using N_T transmit antennas.

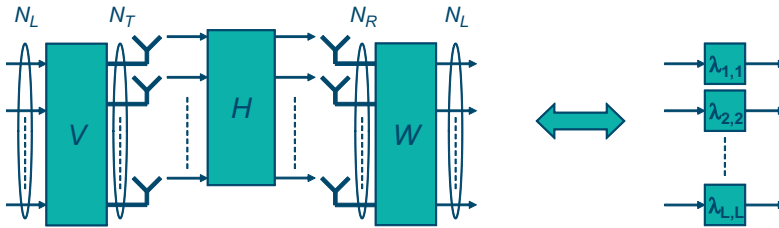
It should be noted that precoder-based spatial multiplexing can be seen as a generalization of precoder-based beam-forming as described in Section 5.4.2, with the precoding vector of size $N_T \times 1$ replaced by a precoding matrix of size $N_T \times N_L$.

The precoding of Figure 5.16 can serve two purposes:

- In the case when the number of signals to be spatially multiplexed equals the number of transmit antennas ($N_L = N_T$), the precoding can be used to “orthogonalize” the parallel transmissions, allowing for improved signal isolation at the receiver side.
- In the case when the number of signals to be spatially multiplexed is less than the number of transmit antennas ($N_L < N_T$), the precoding also provides the mapping of the N_L spatially multiplexed signals to the N_T transmit antennas, including the combination of spatial multiplexing and beam-forming.

To confirm that precoding can improve the isolation between the spatially multiplexed signals, express the channel matrix H as its singular-value decomposition [38]:

$$H = W \cdot \Sigma \cdot V^H, \quad (5.15)$$

**FIGURE 5.17**

Orthogonalization of spatially multiplexed signals by means of precoding. $\lambda_{i,i}$ is the i th eigenvalue of the matrix $H^H H$.

where the columns of V and W each form an orthonormal set and Σ is an $N_L \times N_L$ diagonal matrix with the N_L strongest eigenvalues of $H^H H$ as its diagonal elements. By applying the matrix V as precoding matrix at the transmitter side and the matrix W^H at the receiver side, one arrives at an equivalent channel matrix $H' = \Sigma$ (see Figure 5.17). As H' is a diagonal matrix, there is thus no interference between the spatially multiplexed signals at the receiver. At the same time, as both V and W have orthonormal columns, the transmit power as well as the demodulator noise level (assuming spatially white noise) are unchanged.

Clearly, in the case of precoding each received signal will have a certain “quality”, depending on the eigenvalues of the channel matrix (see right part of Figure 5.17). This indicates potential benefits of applying dynamic link adaptation in the *spatial domain* – that is, the adaptive selection of the coding rates and/or modulation schemes for each signal to be transmitted.

As the precoding matrix will never perfectly match the channel matrix in practice, there will always be some residual interference between the spatially multiplexed signals. This interference can be taken care of by means of additional receiver-side linear processing according to Figure 5.15 or nonlinear processing as discussed in Section 5.5.3 below.

To determine the precoding matrix V , knowledge about the channel matrix H is obviously needed. Similar to precoder-based beam-forming, a common approach is to have the receiver estimate the channel and decide on a suitable precoding matrix from a set of available precoding matrices (the precoder *codebook*). The receiver then feeds back information about the selected precoding matrix to the transmitter.

5.5.3 Nonlinear Receiver Processing

The previous sections discussed the use of linear receiver processing to jointly recover spatially multiplexed signals. However, improved demodulation performance can be achieved if nonlinear receiver processing can be applied in the case of spatial multiplexing.

The “optimal” receiver approach for spatially multiplexed signals is to apply *Maximum-Likelihood* (ML) detection [25]. However, in many cases ML detection is too complex to use. Thus, several different proposals have been made for reduced complexity, almost ML, schemes (see, for example, [39]).

Another nonlinear approach to the demodulation of spatially multiplexed signals is to apply *Successive Interference Cancellation* (SIC) [40]. Successive Interference Cancellation is based on an assumption that the spatially multiplexed signals are separately coded before the spatial multiplexing.

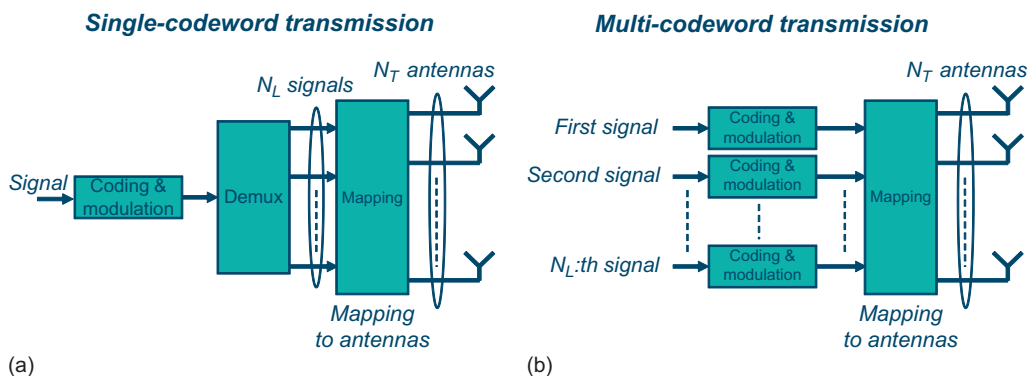


FIGURE 5.18

Single-codeword transmission (a) vs. multi-codeword transmission (b).

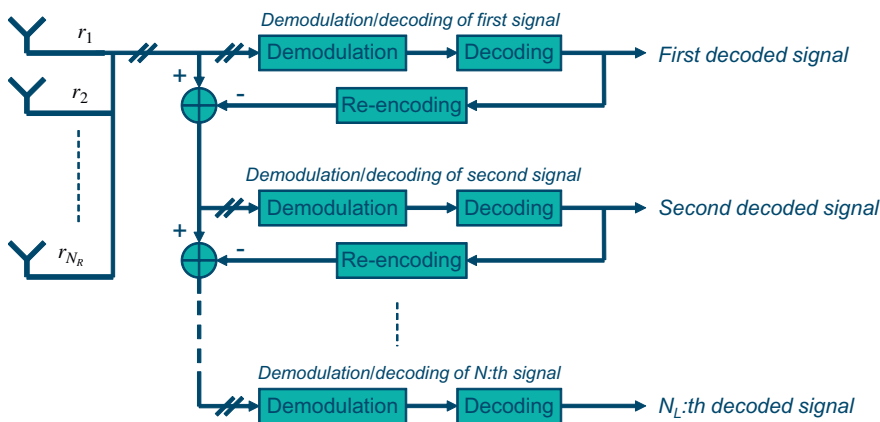


FIGURE 5.19

Demodulation/decoding of spatially multiplexed signals based on Successive Interference Cancellation.

This is often referred to as *Multi-Codeword* transmission, in contrast to *Single-Codeword* transmission where the spatially multiplexed signals are assumed to be jointly coded (Figure 5.18). It should be understood that, also in the case of multi-codeword transmission, the data may originate from the same source but is then demultiplexed into different signals to be spatially multiplexed before channel coding.

As shown in Figure 5.19, in the case of successive interference cancellation the receiver first demodulates and decodes one of the spatially multiplexed signals. The corresponding decoded data is then, if correctly decoded, re-encoded and subtracted from the received signals. A second spatially multiplexed signal can then be demodulated and decoded without, at least in the ideal case, any interference from the first signal – that is, with an improved signal-to-interference ratio. The decoded data of the second signal is then, if correctly decoded, re-encoded and subtracted from the received signal

before decoding of a third signal. These iterations continue until all spatially multiplexed signals have been demodulated and decoded.

Clearly, in the case of Successive Interference Cancellation, the first signals to be decoded are subject to higher interference level, compared to later decoded signals. To work properly, there should thus be a differentiation in the robustness of the different signals with, at least in principle, the first signal to be decoded being more robust than the second signal, the second signal being more robust than the third signal, etc. Assuming multi-codeword transmission according to [Figure 5.18b](#), this can be achieved by applying different modulation schemes and coding rates to the different signals with, typically, lower-order modulation and lower coding rate, implying a lower data rate, for the first signals to be decoded. This is often referred to as *Per-Antenna Rate Control* (PARC) [41].

Scheduling, Link Adaptation, and Hybrid ARQ

One key characteristic of mobile radio communication is the typically rapid and significant variations in the instantaneous channel conditions. There are several reasons for these variations. Frequency-selective fading will result in rapid and random variations in the channel attenuation. Shadow fading and distance-dependent path loss will also significantly affect the average received signal strength. Finally, the interference at the receiver due to transmissions in other cells and by other terminals will also impact the interference level. Hence, to summarize, there will be rapid, and to some extent random, variations in the experienced quality of each radio link in a cell, variations that must be taken into account and preferably exploited.

In this chapter, some of the techniques for handling variations in the instantaneous radio-link quality will be discussed. *Channel-dependent scheduling* in a mobile-communication system deals with the question of how to share, between different users (different terminals), the radio resource(s) available in the system to achieve as efficient resource utilization as possible. Typically, this implies minimizing the amount of resources needed per user and thus allowing for as many users as possible in the system, while still satisfying whatever quality-of-service requirements that may exist. Closely related to scheduling is *link adaptation*, which deals with how to set the transmission parameters of a radio link to handle variations of the radio-link quality.

Both channel-dependent scheduling and link adaptation try to exploit the channel variations through appropriate processing *prior* to transmission of the data. However, due to the random nature of the variations in the radio-link quality, perfect adaptation to the instantaneous radio-link quality is never possible. *Hybrid ARQ*, which requests retransmission of erroneously received data packets, is therefore useful. This can be seen as a mechanism for handling variations in the instantaneous radio-link quality *after* transmission and nicely complements channel-dependent scheduling and link adaptation. Hybrid ARQ also serves the purpose of handling random errors due to, for example, noise in the receiver.

6.1 LINK ADAPTATION: POWER AND RATE CONTROL

Historically, *dynamic transmit-power control* has been used in CDMA-based mobile-communication systems such as WCDMA and CDMA2000 to compensate for variations in the instantaneous channel conditions.¹ As the name suggests, dynamic power control dynamically adjusts the radio-link transmit

¹Power control is also used in GSM, but operating on a much slower basis than in CDMA-based systems. In CDMA-based systems, power control is also essential to handle the near-far problem and the associated performance impact from non-orthogonal uplink transmissions.

power to compensate for variations and differences in the instantaneous channel conditions. The aim of these adjustments is to maintain a (near) constant E_b/N_0 at the receiver to successfully transmit data without the error probability becoming too high. In principle, transmit-power control increases the power at the transmitter when the radio link experiences poor radio conditions (and vice versa). Thus, the transmit power is in essence inversely proportional to the channel quality, as illustrated in Figure 6.1a. This results in a basically constant data rate, regardless of the channel variations. For services such as circuit-switched voice, this is a desirable property. Transmit-power control can be seen as one type of link adaptation – that is, the adjustment of transmission parameters, in this case the transmit power – to adapt to differences and variations in the instantaneous channel conditions to maintain the received E_b/N_0 at a desired level.

However, in many cases of mobile communication, especially in the case of packet-data traffic, there is not a strong need to provide a certain constant data rate over a radio link. Rather, from a user perspective, the data rate provided over the radio interface should simply be as “high as possible”. Actually, even in case of typical “constant-rate” services such as voice and video, (short-term)

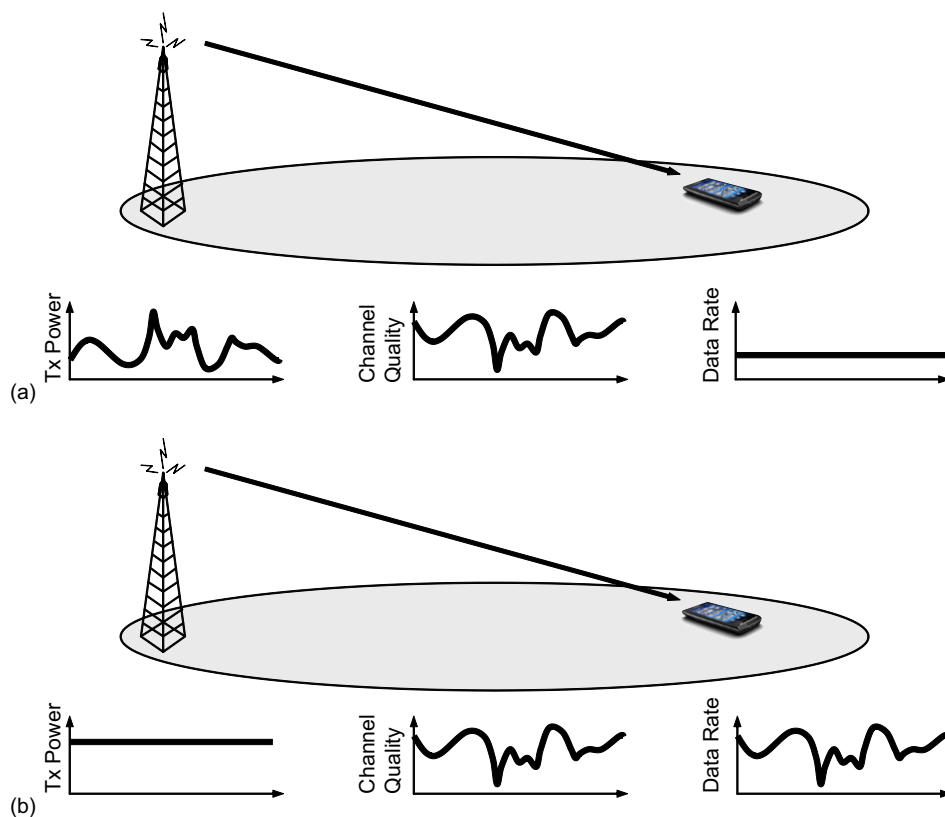


FIGURE 6.1

(a) Power control (top). (b) Rate control (bottom).

variations in the data rate are often not an issue, as long as the average data rate remains constant, assuming averaging over some relatively short time interval. In such cases – that is, when a constant data rate is not required – an alternative to transmit-power control is link adaptation by means of *dynamic rate control*. Rate control does not aim at keeping the instantaneous radio-link data rate constant, regardless of the instantaneous channel conditions. Instead, with rate control, the data rate is dynamically adjusted to compensate for the varying channel conditions. In situations with advantageous channel conditions, the data rate is increased, and vice versa. Thus, rate control maintains the $E_b/N_0 \sim P/R$ at the desired level, not by adjusting the transmission power P , but rather by adjusting the data rate R . This is illustrated in Figure 6.1b.

It can be shown that rate control is more efficient than power control [42,43]. Rate control in principle implies that the power amplifier is always transmitting at full power and therefore efficiently utilized. Power control, on the other hand, results in the power amplifier in most situations not being efficiently utilized as the transmission power is less than its maximum.

In practice, the radio-link data rate is controlled by adjusting the modulation scheme and/or the channel coding rate. In the case of advantageous radio-link conditions, the E_b/N_0 at the receiver is high and the main limitation of the data rate is the bandwidth of the radio link. Hence, in such situations higher-order modulation, for example 16QAM or 64QAM, together with a high code rate, is appropriate as discussed in Chapter 2. Similarly, in the case of poor radio-link conditions, QPSK and low-rate coding is used. For this reason, link adaptation by means of rate control is sometimes also referred to as *Adaptive Modulation and Coding* (AMC).

6.2 CHANNEL-DEPENDENT SCHEDULING

Scheduling controls the allocation of the shared resources among users at each time instant. It is closely related to link adaptation, and often scheduling and link adaptation is seen as one joint function. The scheduling principles, as well as which resources are shared between users, differ depending on the radio-interface characteristics – for example, whether uplink or downlink is considered and whether different users' transmissions are mutually orthogonal or not.

6.2.1 Downlink Scheduling

In the downlink, transmissions to different terminals within a cell are typically mutually orthogonal, implying that, at least in theory, there is no interference between the transmissions (no intra-cell interference). Downlink intra-cell orthogonality can be achieved in the time domain (*Time-Division Multiplexing*, TDM), in the frequency domain (*Frequency-Domain Multiplexing*, FDM), or in the code domain (*Code-Domain Multiplexing*, CDM). In addition, the spatial domain can also be used to separate users, at least in a quasi-orthogonal way, through different antenna arrangements. This is sometimes referred to as *Spatial-Division Multiplexing* (SDM), although in most cases it is used in combination with one or several of the above multiplexing strategies and is not discussed further in this chapter.

For packet data, where the traffic is often very bursty, it can be shown that TDM is preferable from a theoretical point of view [44,45] and is therefore typically the main component in the downlink [46,47]. However, as discussed in Chapter 3, the TDM component is often combined with

sharing of the radio resource also in the frequency domain (FDM) or in the code domain (CDM). For example, in the case of HSPA (see Chapter 19), downlink multiplexing is a combination of TDM and CDM. On the other hand, in the case of LTE, downlink multiplexing is a combination of TDM and FDM. The reasons for sharing the resources not only in the time domain will be elaborated upon later in this section.

When transmissions to multiple users occur in parallel, either by using FDM or CDM, there is also an instantaneous sharing of the total available cell transmit power. In other words, not only are the time/frequency/code resources shared resources, but also the power resource in the base station. In contrast, in the case of sharing only in the time domain there is, by definition, only a single transmission at a time and thus no instantaneous sharing of the total available cell transmit power.

For the purpose of discussion, assume initially a TDM-based downlink with a single user being scheduled at a time. In this case, the utilization of the radio resources is maximized if, at each time instant, all resources are assigned to the user with the best instantaneous channel condition:

- In the case of link adaptation based on power control, this implies that the lowest possible transmit power can be used for a given data rate and thus minimizes the interference to transmissions in other cells for a given link utilization.
- In the case of link adaptation based on rate control, this implies that the highest data rate is achieved for a given transmit power or, in other words, for a given interference to other cells, the highest link utilization is achieved.

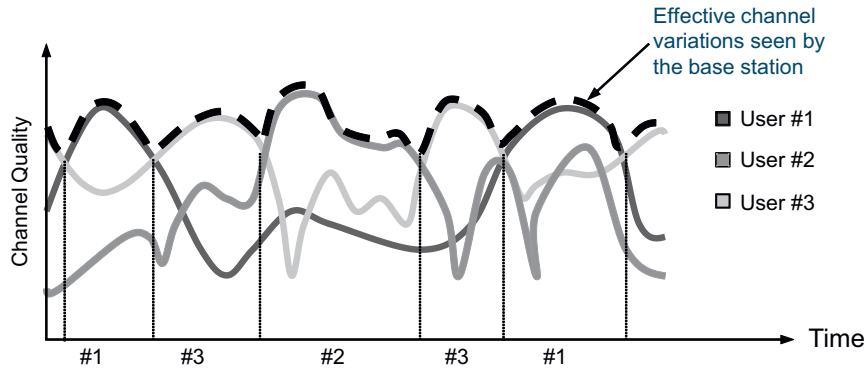
However, if applied to the downlink, transmit-power control in combination with TDM scheduling implies that the total available cell transmit power will, in most cases, not be fully utilized. Thus, rate control is generally preferred [42,44,47,48].

The strategy outlined above is an example of channel-dependent scheduling, where the scheduler takes the instantaneous radio-link conditions into account. Scheduling the user with the instantaneously best radio-link conditions is often referred to as *max-C/I* (or *maximum rate*) scheduling. Since the radio conditions for the different radio links within a cell typically vary independently, at each point in time there is almost always a radio link whose channel quality is near its peak (see Figure 6.2). Thus, the channel eventually used for transmission will typically have a high quality and, with rate control, a correspondingly high data rate can be used. This translates into a high system capacity. The gain obtained by transmitting to users with favorable radio-link conditions is commonly known as multi-user diversity; the multi-user diversity gains are larger, the larger the channel variations and the larger the number of users in a cell. Hence, in contrast to the traditional view that fast fading – that is, rapid variations in the radio-link quality – is an undesirable effect that has to be combated, the possibility of channel-dependent scheduling implies that *fading is in fact potentially beneficial and should be exploited*.

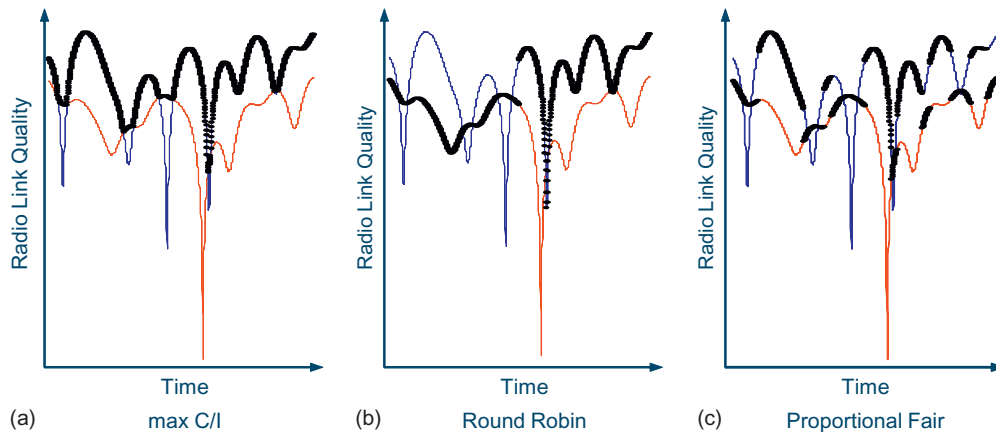
Mathematically, the max-C/I (maximum rate) scheduler can be expressed as scheduling user k given by:

$$k = \arg \max_i R_i,$$

where R_i is the instantaneous data rate for user i . Although, from a system capacity perspective, max-C/I scheduling is beneficial, this scheduling principle will not be fair in all situations. If all terminals are, on average, experiencing similar channel conditions and large variations in the instantaneous channel conditions are only due to, for example, fast multi-path fading, all users will experience the

**FIGURE 6.2**

Channel-dependent scheduling.

**FIGURE 6.3**

Examples of three different scheduling behaviors for two users with different average channel quality: (a) max-C/I; (b) round robin; (c) proportional fair. The selected user is shown with bold lines.

same average data rate. Any variations in the instantaneous data rate are rapid and often not even noticeable by the user. However, in practice different terminals will also experience differences in the (short-term) average channel conditions – for example, due to differences in the distance and shadow fading between the base station and the terminal. In this case, the channel conditions experienced by one terminal may, for a relatively long time, be worse than the channel conditions experienced by other terminals. A pure max-C/I-scheduling strategy may then, in essence, “starve” the terminal with the bad channel conditions, and the terminal with bad channel conditions will never be scheduled. This is illustrated in Figure 6.3a, where a max-C/I scheduler is used to schedule between two different users with different average channel quality. The same user is scheduled virtually all the

time. Although resulting in the highest system capacity, this situation is often not acceptable from a quality-of-service point of view.

An alternative to the max-C/I scheduling strategy is so-called *round-robin* scheduling, illustrated in Figure 6.3b. This scheduling strategy lets users take turns in using the shared resources, without taking the instantaneous channel conditions into account. Round-robin scheduling can be seen as fair scheduling in the sense that the same amount of radio resources (the same amount of time) is given to each communication link. However, round-robin scheduling is not fair in the sense of providing the same service quality to all communication links. In that case more radio resources (more time) must be given to communication links with bad channel conditions. Furthermore, as round-robin scheduling does not take the instantaneous channel conditions into account in the scheduling process, it will lead to lower overall system performance but more equal service quality between different communication links, compared to max-C/I scheduling.

Thus, what is needed is a scheduling strategy that is able to utilize the fast channel variations to improve the overall cell throughput while still ensuring the same average user throughput for all users or at least a certain minimum user throughput for all users. When discussing and comparing different scheduling algorithms it is important to distinguish between different types of variations in the service quality:

- Fast variations in the service quality corresponding to, for example, fast multi-path fading and fast variations in the interference level. For many packet-data applications, relatively large short-term variations in service quality are often acceptable or not even noticeable to the user.
- More long-term differences in the service quality between different communication links corresponding to, for example, differences in the distance to the cell site and shadow fading. In many cases there is a need to limit such long-term differences in service quality.

A practical scheduler should thus operate somewhere in between the max-C/I scheduler and the round-robin scheduler – that is, try to utilize fast variations in channel conditions as much as possible while still satisfying some degree of fairness between users.

One example of such a scheduler is the proportional-fair scheduler [49–51], illustrated in Figure 6.3c. In this strategy, the shared resources are assigned to the user with the *relatively* best radio-link conditions – that is, at each time instant, user k is selected for transmission according to:

$$k = \arg \max_i \frac{R_i}{\bar{R}_i},$$

where R_i is the instantaneous data rate for user i and $\bar{R}_i$ is the average data rate for user i . The average is calculated over a certain averaging period T_{PF} . To ensure efficient usage of the short-term channel variations and, at the same time, limit the long-term differences in service quality to an acceptable level, the time constant T_{PF} should be set longer than the time constant for the short-term variations. At the same time, T_{PF} should be sufficiently short so that quality variations within the interval T_{PF} are not strongly noticed by a user. Typically, T_{PF} can be set to be of the order of one second.

In the above discussion, it was assumed that all the radio resources in the downlink were assigned to a single user at a time – that is, scheduling was done purely in the time domain using TDM between users. However, in several situations, TDM is complemented by CDM or FDM. In principle, there are two reasons for not relying solely on TDM in the downlink:

- In the case of insufficient payload – that is, the amount of data to transfer to a user is not sufficiently large to utilize the full channel capacity, and a fraction of resources could be assigned to another user, either through FDM or CDM.
- In the case where channel variations in the frequency domain are exploited through FDM, as discussed further below.

The scheduling strategies in these cases can be seen as generalizations of the schemes discussed for the TDM-only cases above. For example, to handle small payloads, a greedy filling approach can be used, where the scheduled user is selected according to max-C/I (or any other scheduling scheme). Once this user has been assigned resources matching the amount of data awaiting transmission, the second best user according to the scheduling strategy is selected and assigned (a fraction of) the residual resources and so on.

Finally, it should also be noted that the scheduling algorithm typically is a base-station-implementation issue and nothing that is normally specified in any standard. What need to be specified in a standard to support channel-dependent scheduling are channel-quality measurements/reports and the signaling needed for dynamic resource allocation.

6.2.2 Uplink Scheduling

The previous section discussed scheduling from a downlink perspective. However, scheduling is equally applicable to uplink transmissions and to a large extent the same principles can be considered, although there are some differences between the two.

Fundamentally, the uplink power resource is *distributed* among the users, while in the downlink the power resource is *centralized* within the base station. Furthermore, the maximum uplink transmission power of a single terminal is typically significantly lower than the output power of a base station. This has a significant impact on the scheduling strategy. Unlike the downlink, where pure TDMA can often be used, uplink scheduling typically has to rely on sharing in the frequency and/or code domain in addition to the time domain, as a single terminal may not have sufficient power for efficiently utilizing the link capacity.

Similar to the downlink case, channel-dependent scheduling is also beneficial in the uplink case. However, the characteristics of the underlying radio interface, most notably whether the uplink relies on orthogonal or non-orthogonal multiple access and the type of link adaptation scheme used, also have a significant impact on the uplink scheduling strategy.

In the case of a non-orthogonal multiple-access scheme such as CDMA, power control is typically essential for proper operation. As discussed earlier in this chapter, the purpose of power control is to control the received E_b/N_0 such that the received information can be recovered. However, in a non-orthogonal multiple-access setting, power control also serves the purpose of controlling the amount of interference affecting *other* users. This can be expressed as the *maximum tolerable interference level* at the base station in a shared resource. Even if, from a single user's perspective, it would be beneficial to transmit at full power to maximize the data rate, this may not be acceptable from an interference perspective as other terminals in this case may not be able to successfully transfer any data. Thus, with non-orthogonal multiple access, scheduling a terminal when the channel conditions are favorable may not directly translate into a higher data rate as the interference generated to other simultaneously transmitting terminals in the cell must be taken into account. Stated differently, the

received power (and thus the data rate) is, thanks to power control, in principle constant, regardless of the channel conditions at the time of transmission, while the *transmitted* power depends on the channel conditions at the time of transmission. Hence, even though channel-dependent scheduling in this example does not give a direct gain in terms of a higher data rate from the terminal, channel-dependent scheduling will still provide a gain for the system in terms of lower intra-cell interference.

The above discussion on non-orthogonal multiple access was simplified in the sense that no bounds on the terminals' transmission power were assumed. In practice, the transmission power of a terminal is upper bounded, due both to implementation and regulatory reasons, and scheduling a terminal for transmission in favorable channel conditions decreases the probability that the terminal has insufficient power to utilize the channel capacity.

In the case of an orthogonal multiple-access scheme, intra-cell power control is fundamentally not necessary and the benefits of channel-dependent scheduling become more similar to the downlink case. In principle, from an intra-cell perspective, a terminal can transmit at full power and the scheduler assigns a suitable part of the orthogonal resources (in practice a suitable part of the overall bandwidth) to the terminal for transmission. The remaining orthogonal resources can be assigned to other users. However, implementation constraints, for example leakage between the received signals or limited dynamic range in the receiver circuitry, may pose restrictions on the maximum tolerable power difference between the signals from simultaneously transmitting terminals. As a consequence, a certain degree of power control may be necessary, making the situation somewhat similar to the non-orthogonal case.

The discussion on non-orthogonal and orthogonal multiple access mainly considered intra-cell multiple access. However, in many practical systems universal frequency reuse between cells is applied. In this case, the inter-cell multiple access is non-orthogonal, regardless of the intra-cell multiple access, which sets limits on the allowable transmission power from a terminal.

Regardless of whether orthogonal or non-orthogonal multiple access is used, the same basic scheduling principles as for the downlink can be used. A max-C/I scheduler would assign all the uplink resources to the terminal with the best uplink channel conditions. Neglecting any power limitations in the terminal, this would result in the highest capacity (in an isolated cell) [44].

In the case of a non-orthogonal multiple-access scheme, *greedy filling* is one possible scheduling strategy [48]. With greedy filling, the terminal with the best radio conditions is assigned as high a data rate as possible. If the interference level at the receiver is smaller than the maximum tolerable level, the terminal with the second best channel conditions is allowed to transmit as well, continuing with more and more terminals until the maximum tolerable interference level at the receiver is reached. This strategy maximizes the air interface utilization but is achieved at the cost of potentially large differences in data rates between users. In the extreme case, a user at the cell border with poor channel conditions may not be allowed to transmit at all.

Strategies between greedy filling and max-C/I can also be envisioned – for example, different proportional-fair strategies. This can be achieved by including a weighting factor for each user, proportional to the ratio between the instantaneous and average data rates, into the greedy filling algorithm.

The schedulers above all assume knowledge of the instantaneous radio-link conditions, knowledge that can be hard to obtain in the uplink scenario, as discussed in Section 7.2.4. In situations when no information about the uplink radio-link quality is available at the scheduler, *round-robin* scheduling can be used. Similar to the downlink, round-robin implies terminals taking turns in transmitting, thus creating a TDMA-like operation with inter-user orthogonality in the time domain. Although the round-robin scheduler is simple, it is far from the optimal scheduling strategy.

However, as already discussed in Chapter 4, the transmission power in a terminal is limited and therefore additional sharing of the uplink resources in the frequency and/or code domain is required. This also impacts the scheduling decisions. For example, terminals far from the base station typically operate in the power-limited region, in contrast to terminals close to the base stations, which often are in the bandwidth-limited region (for a discussion on power-limited vs. bandwidth-limited operation, see Chapter 2). Thus, for a terminal far from the base station, increasing the bandwidth will not result in an increased data rate and it is better to only assign a small amount of the bandwidth to this terminal and assign the remaining bandwidth to other terminals. On the other hand, for terminals close to the base station, an increase in the assigned bandwidth will provide a higher data rate.

6.2.3 Link Adaptation and Channel-Dependent Scheduling in the Frequency Domain

In the previous section, TDM-based scheduling was assumed and it was explained how, in this case, channel variations in the time domain could be utilized to improve system performance by applying channel-dependent scheduling, especially in combination with dynamic rate control. However, if the scheduler has access to the frequency domain, for example through the use of OFDM transmission, scheduling and link adaptation can also take place in the frequency domain.

Link adaptation in the frequency domain implies that, based on knowledge about the instantaneous channel conditions also in the frequency domain – that is, knowledge about the attenuation as well as the noise/interference level of, in the extreme case, every OFDM subcarrier – the power and/or the data rate of each OFDM carrier can be individually adjusted for optimal utilization.

Similarly, channel-dependent scheduling in the frequency domain implies that, based on knowledge about the instantaneous channel conditions also in the frequency domain, different subcarriers are used for transmission to or from different terminals. The scheduling gains from exploiting variations in the frequency domain are similar to those obtained from time-domain variations. Obviously, in situations where the channel quality varies significantly with the frequency while the channel quality only varies slowly with time, channel-dependent scheduling in the frequency domain can enhance system capacity. An example of such a situation is a wideband indoor system with low mobility, where the quality only varies slowly with time.

6.2.4 Acquiring on Channel-State Information

To select a suitable data rate, in practice a suitable modulation scheme and channel-coding rate, the transmitter needs information about the radio-link channel conditions. Such information is also required for the purpose of channel-dependent scheduling. In the case of a system based on frequency-division duplex (FDD), only the receiver can accurately estimate the radio-link channel conditions.

For the downlink, most systems provide a downlink signal of a predetermined structure, known as the downlink pilot or the downlink reference signal. This reference signal is transmitted from the base station with constant power and can be used by the terminal to estimate the instantaneous downlink channel conditions. Information about the instantaneous downlink conditions can then be reported to the base station.

Basically, what is relevant for the transmitter is an estimate reflecting the channel conditions *at the time of transmission*. Hence, in principle, the terminal could apply a predictor, trying to predict the future channel conditions and report this predicted value to the base station. However, as this would

require specification of prediction algorithms and how they would operate when the terminal is moving at different speeds, most practical systems simply report the measured channel conditions to the base station. This can be seen as a very simple predictor, basically assuming the conditions in the near future will be similar to the current conditions. Thus, the more rapid the time-domain channel variations, the less efficient the link adaptation.

As there inevitably will be a delay between the point in time when the terminal measured the channel conditions and the application of the reported value in the transmitter, channel-dependent scheduling and link adaptation typically operates at its best at low terminal mobility. If the terminal starts to move at high speed, the measurement reports will be outdated once arriving at the base station. In such cases, it is often preferable to perform link adaptation on the long-term average channel quality and rely on hybrid ARQ with soft combining for the rapid adaptation.

For the uplink, estimation of the uplink channel conditions is not as straightforward, as there is typically no reference signal transmitted with constant power from each terminal. Means to estimate uplink channel conditions for LTE are discussed in Chapter 11.

In the case of a system with time-division duplex (TDD), where uplink and downlink communication are time multiplexed within the same frequency band, the instantaneous uplink signal quality attenuation could be estimated from downlink measurements of the terminal, due to the reciprocity of the multi-path fading in the case of TDD. However, it should then be noted that this may not provide full knowledge of the downlink channel conditions. As an example, the interference situations at the terminal and the base station are different also in the case of TDD.

6.2.5 Traffic Behavior and Scheduling

It should be noted that there is little difference between different scheduling algorithms at low system load – that is, when only one or, in some cases, a few users have data waiting for transmission at the base station at each scheduling instant. The differences between different scheduling algorithms are primarily visible at high load. However, not only the load, but also the traffic behavior affects the overall scheduling performance.

As discussed above, channel-dependent scheduling tries to exploit short-term variations in radio quality. Generally, a certain degree of long-term fairness in service quality is desirable, which should be accounted for in the scheduler design. However, since system throughput decreases the more fairness is enforced, a trade-off between fairness and system throughput is necessary. In this trade-off, it is important to take traffic characteristics into account as they have a significant influence on the trade-off between system throughput and service quality.

To illustrate this, consider three different downlink schedulers:

1. *Round-robin (RR) scheduler*, where channel conditions are not taken into account.
2. *Proportional-fair (PF) scheduler*, where short-term channel variations are exploited while maintaining the long-term average user data rate.
3. *Max-C/I scheduler*, where the user with the best instantaneous channel quality in absolute terms is scheduled.

For a full buffer scenario when there is always data available at the base station for all terminals in the cell, a max-C/I scheduler will result in no, or very low, user throughput for users at the cell edge with a low average channel quality. The reason is the fundamental strategy of the max-C/I

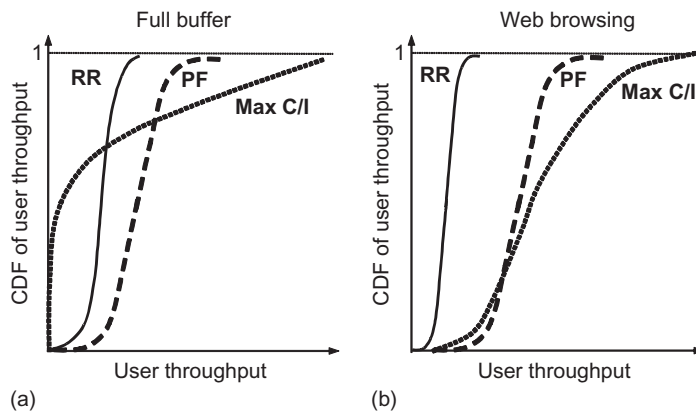
**FIGURE 6.4**

Illustration of the principle behavior of different scheduling strategies:
(a) for full buffers; (b) for web-browsing traffic model.

scheduler – all resources are allocated for transmission to the terminal whose channel conditions support the highest data rate. Only in the rare, not to say unlikely, case of a cell-edge user having better conditions than a cell-center user, for example due to a deep fading dip for the cell-center user, will the cell-edge user be scheduled. A proportional-fair scheduler, on the other hand, will ensure some degree of fairness by selecting the user supporting the highest data rate relative to its average data rate. Hence, users tend to be scheduled on their fading peaks, regardless of the absolute quality. Thus, users on the cell edge will also be scheduled, thereby resulting in some degree of fairness between users.

For a scenario with bursty packet data, the situation is different. In this case, the users' buffers will be finite and in many cases also empty. For example, a web page has a certain size and, after transmitting the page, there is no more data to be transmitted to the terminal in question until the users requests a new page by clicking on a link. In this case, a max-C/I scheduler can still provide a certain degree of fairness. Once the buffer for the user with the highest C/I has been emptied, another user with non-empty buffers will have the highest C/I and be scheduled and so on. This is the reason for the difference between full buffer and web-browsing traffic illustrated in Figure 6.4. The proportional-fair scheduler has similar performance in both scenarios.

Clearly, the degree of fairness introduced by the traffic properties depends heavily on the actual traffic; a design made with certain assumptions may be less desirable in an actual network where the traffic pattern may be different from the assumptions made during the design. Therefore, relying solely on the traffic properties for fairness is not a good strategy, but the discussion above emphasizes the need to design the scheduler not only for the full buffer case.

6.3 ADVANCED RETRANSMISSION SCHEMES

Transmissions over wireless channels are subject to errors, for example due to variations in the received signal quality. To some degree, such variations can be counteracted through link adaptation,

as discussed above. However, receiver noise and unpredictable interference variations cannot be counteracted. Therefore, virtually all wireless communications systems employ some form of *Forward Error Correction* (FEC), tracing its roots to the pioneering work of Claude Shannon in 1948 [11]. There is a rich literature in the area of error-correction coding (see, for example, [52,53] and references therein), and a detailed description is beyond the scope of this book. In short, the basic principle beyond forward error-correction coding is to introduce redundancy in the transmitted signal. This is achieved by adding *parity bits* to the information bits prior to transmission (alternatively, the transmission could consist of parity bits alone, depending on the coding scheme used). The parity bits are computed from the information bits using a method given by the coding structure used. Thus, the number of bits transmitted over the channel is larger than the number of original information bits and a certain amount of *redundancy* has been introduced in the transmitted signal.

Another approach to handle transmission errors is to use *Automatic Repeat Request* (ARQ). In an ARQ scheme, the receiver uses an error-detection code, typically a Cyclic Redundancy Check (CRC), to detect if the received packet is in error or not. If no error is detected in the received data packet, the received data is declared error-free and the transmitter is notified by sending a positive acknowledgement (ACK). On the other hand, if an error is detected, the receiver discards the received data and notifies the transmitter via a return channel by sending a negative acknowledgement (NAK). In response to an NAK, the transmitter retransmits the same information.

Virtually all modern communication systems, including LTE, employ a combination of forward error-correction coding and ARQ, known as *hybrid ARQ*. Hybrid ARQ uses forward error-correction codes to correct a subset of all errors and relies on error detection to detect uncorrectable errors. Erroneously received packets are discarded and the receiver requests retransmissions of corrupted packets. Thus, it is a combination of FEC and ARQ as described above. Hybrid ARQ was first proposed in [54] and numerous publications on hybrid ARQ have appeared since (see [52] and references therein). Most practical hybrid ARQ schemes are built around a CRC code for error detection and convolutional or Turbo codes for error correction, but in principle any error-detection and error-correction code can be used.

6.4 HYBRID ARQ WITH SOFT COMBINING

The hybrid-ARQ operation described above discards erroneously received packets and requests retransmission. However, despite it not being possible to decode the packet, the received signal still contains information, which is lost by discarding erroneously received packets. This shortcoming is addressed by *hybrid ARQ with soft combining*. In hybrid ARQ with soft combining, the erroneously received packet is stored in a buffer memory and later combined with the retransmission to obtain a single, combined packet that is more reliable than its constituents. Decoding of the error-correction code operates on the combined signal. If the decoding fails (typically a CRC code is used to detect this event), a retransmission is requested.

Retransmission in any hybrid-ARQ scheme must, by definition, represent the same set of information bits as the original transmission. However, the set of coded bits transmitted in each retransmission may be selected differently as long as they represent the same set of information bits. Hybrid ARQ with soft combining is therefore usually categorized into *Chase combining* and *incremental redundancy*, depending on whether the retransmitted bits are required to be identical to the original transmission or not.

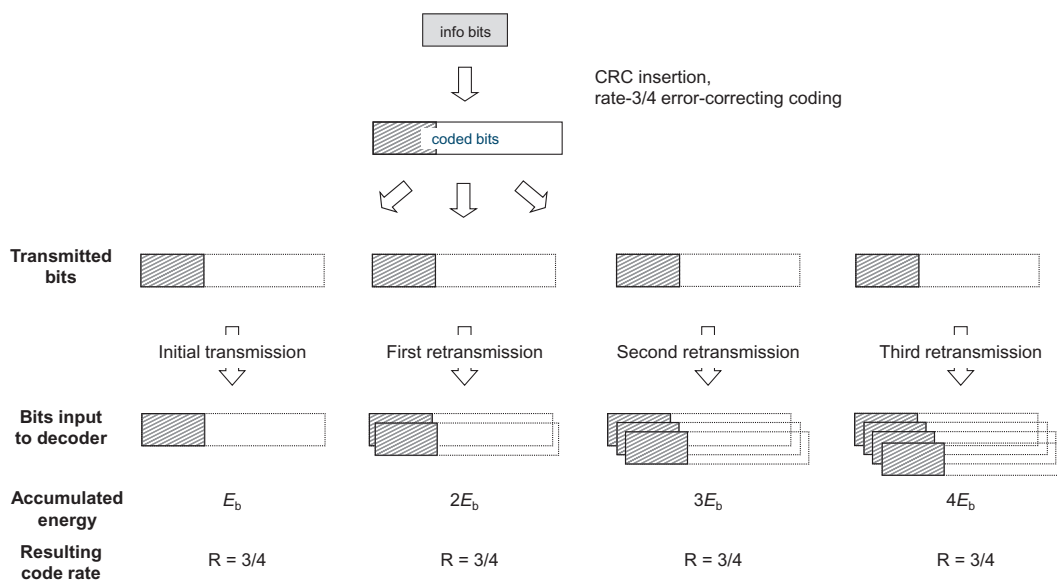


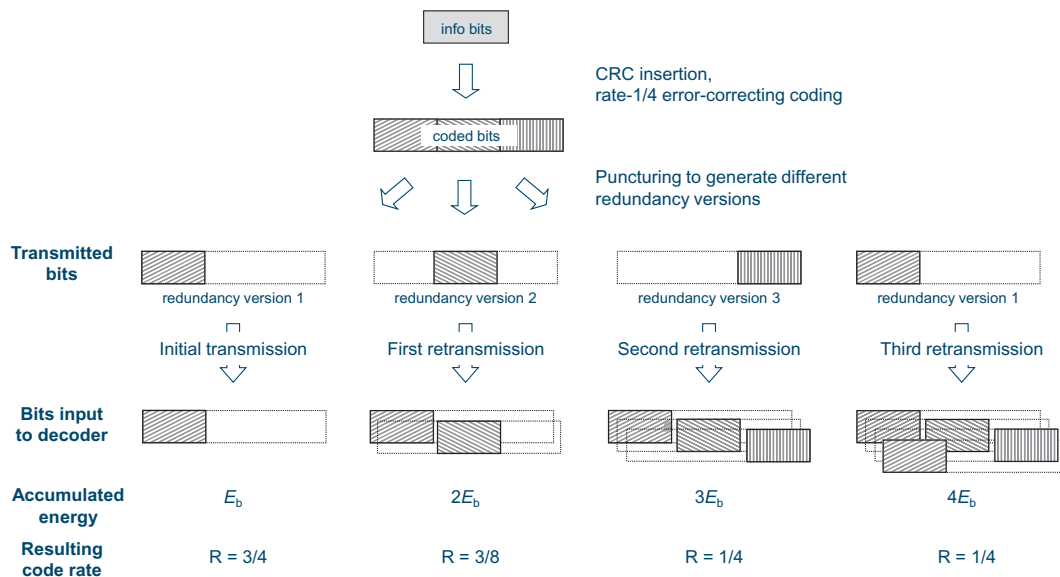
FIGURE 6.5

Example of Chase combining.

Chase combining, where the retransmissions consist of the same set of coded bits as the original transmission, was first proposed in [55]. After each retransmission, the receiver uses maximum-ratio combination to combine each received channel bit with any previous transmissions of the same bit, and the combined signal is fed to the decoder. As each retransmission is an identical copy of the original transmission, retransmissions with Chase combining can be seen as additional repetition coding. Therefore, as no new redundancy is transmitted, Chase combining does not give any additional coding gain but only increases the accumulated received E_b/N_0 for each retransmission (Figure 6.5).

Several variants of Chase combining exist. For example, only a subset of the bits transmitted in the original transmission might be retransmitted, so-called partial Chase combining. Furthermore, although combination is often done after demodulation but before channel decoding, combination can also be carried out at the modulation symbol level before demodulation, as long as the modulation scheme is unchanged between transmission and retransmission.

With *Incremental Redundancy* (IR), each retransmission does not have to be identical to the original transmission. Instead, *multiple sets* of coded bits are generated, each representing the same set of information bits [56,57]. Whenever a retransmission is required, the retransmission typically uses a different set of coded bits than the previous transmission. The receiver combines the retransmission with previous transmission attempts of the same packet. As the retransmission may contain additional parity bits not included in the previous transmission attempts, the resulting code rate is generally lowered by a retransmission. Furthermore, each retransmission does not necessarily have to consist of the same number of coded bits as the original and, in general, the modulation scheme can also be different for different

**FIGURE 6.6**

Example of incremental redundancy.

retransmissions. Hence, incremental redundancy can be seen as a generalization of Chase combining or, stated differently, Chase combining is a special case of incremental redundancy.

Typically, incremental redundancy is based on a low-rate code and the different redundancy versions are generated by puncturing the output of the encoder. In the first transmission only a limited number of coded bits are transmitted, effectively leading to a high-rate code. In the retransmissions, additional coded bits are transmitted. As an example, assume a basic rate-1/4 code. In the first transmission, only every third coded bit is transmitted, effectively giving a rate-3/4 code, as illustrated in Figure 6.6. In the case of a decoding error and a subsequent request for a retransmission, additional bits are transmitted, effectively leading to a rate-3/8 code. After a second retransmission the code rate is 1/4. In the case of more than two retransmissions, transmitted coded bits would already be repeated. In addition to a gain in accumulated received E_b/N_0 , incremental redundancy also results in a coding gain for each retransmission. The gain with IR compared to Chase is larger for high initial code rates, while at lower initial coding rates Chase combining is almost as good as IR [58]. Furthermore, as shown in [59], the performance gain of incremental redundancy compared to Chase combining can also depend on the relative power difference between the transmission attempts.

With incremental redundancy, the code used for the first transmission should provide good performance not only when used alone, but also when used in combination with the code for the second transmission. The same holds for subsequent retransmissions. Thus, as the different redundancy versions typically are generated through puncturing of a low-rate mother code, the puncturing patterns should be defined such that all the code bits used by a high-rate code should also be part of any lower-rate codes. In other words, the resulting code rate R_i after transmission attempt i , consisting of

the coded bits from redundancy versions RV_k , $k = 1, \dots, i$, should have similar performance as a good code designed directly for rate R_i . Examples of this for convolutional codes are the so-called rate-compatible convolutional codes [60].

In the discussion so far, it has been assumed that the receiver has received all the previously transmitted redundancy versions. If all redundancy versions provide the same amount of information about the data packet, the order of the redundancy versions is not critical. However, for some code structures, not all redundancy versions are of equal importance. One example here is Turbo codes, where the systematic bits are of higher importance than the parity bits. Hence, the initial transmission should at least include all the systematic bits and some parity bits. In the retransmission(s), parity bits not in the initial transmission can be included. However, if the initial transmission was received with poor quality or not at all, a retransmission with only parity bits is not appropriate as a retransmission of (some of) the systematic bits provides better performance. Incremental redundancy with Turbo codes can therefore benefit from multiple levels of feedback, for example by using two different negative acknowledgements – NAK to request additional parity bits and LOST to request a retransmission of the systematic bits. In general, the problem of determining the amount of systematic and parity bits in a retransmission based on the signal quality of previous transmission attempts is non-trivial.

Hybrid ARQ with soft combining, regardless of whether Chase or incremental redundancy is used, leads to an implicit reduction of the data rate by means of retransmissions and can thus be seen as *implicit* link adaptation. However, in contrast to link adaptation based on explicit estimates of the instantaneous channel conditions, hybrid ARQ with soft combining implicitly adjusts the coding rate based on the result of the decoding. In terms of overall throughput this kind of implicit link adaptation can be superior to explicit link adaptation, as additional redundancy is only added *when needed* – that is, when previous higher-rate transmissions were not possible to decode correctly. Furthermore, as it does not try to predict any channel variations, it works equally well, regardless of the speed at which the terminal is moving. Since implicit link adaptation can provide a gain in system throughput, a valid question is why explicit link adaptation is necessary at all. One major reason for having explicit link adaptation is the reduced delay. Although relying on implicit link adaptation alone is sufficient from a system throughput perspective, the end-user service quality may not be acceptable from a delay perspective.

LTE Radio Access: An Overview

The work on LTE was initiated in late 2004 with the overall aim of providing a new radio-access technology focusing on packet-switched data only. The first phase of the 3GPP work on LTE was to define a set of performance and capability targets for LTE [10]. This included targets on peak data rates, user/system throughput, spectral efficiency, and control/user-plane latency. In addition, requirements were also set on spectrum flexibility, as well as on interaction/compatibility with other 3GPP radio-access technologies (GSM, WCDMA/HSPA, and TD-SCDMA).

Once the targets were set, 3GPP studies on the feasibility of different technical solutions considered for LTE were followed by development of the detailed specifications. The first release of the LTE specifications, release 8, was completed in the spring of 2008 and commercial network operation began in late 2009. Release 8 has then been followed by additional LTE releases, introducing additional functionality and capabilities in different areas, as illustrated in [Figure 7.1](#).

In parallel to the development of LTE, there has also been an evolution of the overall 3GPP network architecture, termed *System Architecture Evolution* (SAE), including both the radio-access network and the core network. Requirements were also set on the architecture evolution, leading to a new flat radio-access-network architecture with a single type of node, the *eNodeB*¹, as well as a new core-network architecture. An excellent description of the LTE-associated core-network architecture, the *Evolved Packet Core* (EPC), can be found in [9].

The remaining part of this chapter provides an overview of LTE up to and including release 10. The most important technologies used by LTE release 8 – including transmission schemes, scheduling, multi-antenna support, and spectrum flexibility – are covered, as well as the additional features introduced in LTE releases 9 and 10. The chapter can either be read on its own to get a high-level overview of LTE, or as an introduction to the subsequent chapters.

The following chapters, Chapters 8–18, then contain a detailed description of the LTE radio-access technology. Chapter 8 provides an overview of the LTE protocol structure, including RLC, MAC, and the physical layer, explaining the logical and physical channels, and the related data flow. The time–frequency structure on which LTE is based is covered in Chapter 9, followed by a detailed description of the physical layer for downlink and uplink transmission in Chapters 10 and 11 respectively. Chapter 12 contains a description of the retransmission mechanisms used in LTE, followed by a discussion on power control, scheduling, and interference management in Chapter 13. Access procedures, necessary for a terminal to connect to the network, are the topic of Chapter 14. Chapter 15 covers the multi-cast/broadcast functionality of LTE and Chapter 16 describes relaying operation. Chapter 17 addresses how radio-frequency (RF) requirements are defined in LTE, taking into account the spectrum flexibility. Finally, Chapter 18 contains an assessment of the system performance of LTE.

¹ eNodeB is a 3GPP term that can roughly be seen as being equivalent to a base station, see further in Chapter 8.

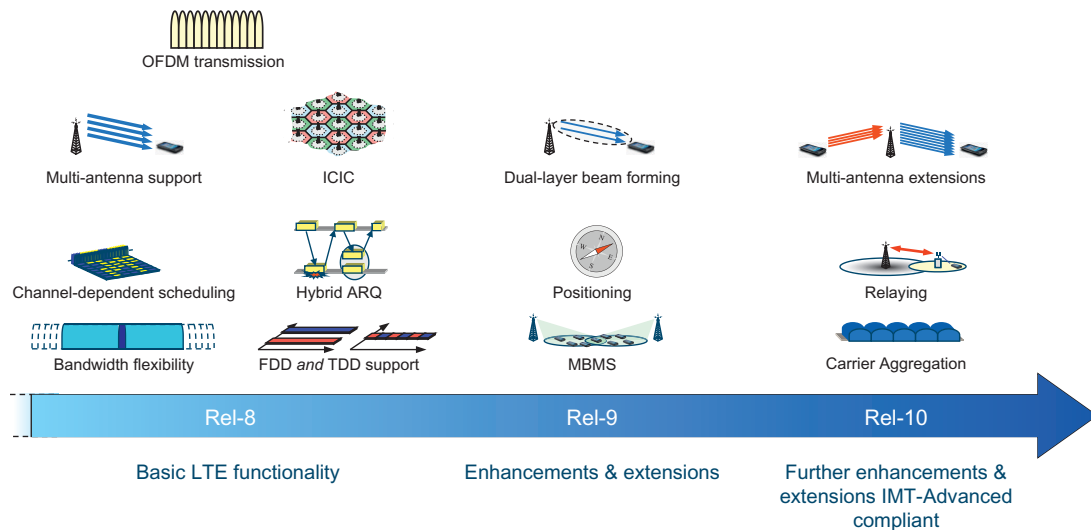


FIGURE 7.1

LTE and its evolution.

7.1 BASIC PRINCIPLES

Building upon the basic technologies described in the previous chapters, the main principles behind LTE will be described in the following.

7.1.1 Transmission Scheme

The LTE downlink transmission scheme is based on conventional OFDM. As discussed in Chapter 3, OFDM is an attractive transmission scheme for several reasons. Due to the relatively long OFDM symbol time in combination with a cyclic prefix, OFDM provides a high degree of robustness against channel frequency selectivity. Although signal corruption due to a frequency-selective channel can, in principle, be handled by equalization at the receiver side, the complexity of such equalization starts to become unattractively high for implementation in a terminal at larger bandwidths and especially in combination with advanced multi-antenna transmission schemes such as spatial multiplexing, thereby making OFDM an attractive choice for LTE for which a wide bandwidth and support for advanced multi-antenna transmission were key requirements.

OFDM also provides some additional benefits relevant for LTE:

- OFDM provides access to the frequency domain, thereby enabling an additional degree of freedom to the channel-dependent scheduler compared to time-domain-only scheduling used in major 3G systems.
- Flexible transmission bandwidth to support operation in spectrum allocations of different size is straightforward with OFDM, at least from a baseband perspective, by varying the number of OFDM subcarriers used for transmission. Note, however, that support of a flexible transmission

bandwidth also requires flexible RF filtering, etc., for which the exact transmission scheme is irrelevant. Nevertheless, maintaining the same baseband-processing structure, regardless of the bandwidth, eases terminal development and implementation.

- Broadcast/multicast transmission, where the same information is transmitted from multiple base stations, is straightforward with OFDM as already described in Chapter 3 and elaborated upon from an LTE perspective in Chapter 15.

Also, the LTE uplink is based on OFDM transmission. However, different means are taken to reduce the *cubic metric* of the uplink transmission, thereby enabling higher terminal power-amplifier efficiency. More specifically, for uplink data transmission, the OFDM modulator is preceded by a DFT precoder, leading to *DFT-spread OFDM* (DFTS-OFDM) as described in Chapter 4. Often, the term DFTS-OFDM is used to describe the LTE uplink transmission scheme in general. However, it should be understood that DFTS-OFDM is only applicable to uplink data transmission. Other means are used to achieve a low cubic metric for other types of uplink transmissions. Thus, the LTE uplink transmission scheme should be described as OFDM with different techniques, including DFT precoding for data transmission, being used to reduce the cubic metric of the transmitted signal.

The use of DFTS-OFDM on the LTE uplink allows for orthogonal separation of uplink transmissions also in the frequency domain. Orthogonal separation is in many cases beneficial as it avoids interference between uplink transmissions from different terminals within the cell (*intra-cell interference*). As discussed in Chapter 4, allocating a very large instantaneous bandwidth for transmission from a single terminal is not an efficient strategy in situations where the data rate is mainly limited by the available terminal transmit power rather than the bandwidth. In such situations, a terminal can instead be allocated only a part of the total available bandwidth and other terminals within the cell can be scheduled to transmit in parallel on the remaining part of the spectrum. In other words, the LTE uplink transmission scheme allows for both *time division* (TDMA) and *frequency division* (FDMA) between users.

7.1.2 Channel-Dependent Scheduling and Rate Adaptation

At the core of the LTE transmission scheme is the use of *shared-channel transmission* with the overall time–frequency resource dynamically shared between users. The use of shared-channel transmission is well matched to the rapidly varying resource requirements posed by packet-data communication and also enables several of the other key technologies on which LTE is based.

The scheduler controls, for each time instant, to which users the different parts of the shared resource should be assigned. The scheduler also determines the data rate to be used for each transmission. Thus, *rate adaptation* can be seen as a part of the scheduling functionality. The scheduler is thus a key element and to a large extent determines the overall system performance, especially in a highly loaded network. Both downlink and uplink transmissions are subject to tight scheduling in LTE. From Chapter 6 it is well known that a substantial gain in system capacity can be achieved if the channel conditions are taken into account in the scheduling decision, so-called *channel-dependent scheduling*. Due to the use of OFDM in both the downlink and uplink transmission directions, the scheduler has access to both the time and frequency domains. In other words, the scheduler can, for each time instant and frequency region, select the user with the best channel conditions, as illustrated in Figure 7.2.