Efficient Resource Allocation and Spectrum Utilisation in Licensed Shared Access Systems

Ntougias, Konstantinos

Publication date:
2020

Document Version
Publisher's PDF, also known as Version of record

EFFICIENT RESOURCE ALLOCATION AND SPECTRUM UTILISATION IN LICENSED SHARED ACCESS SYSTEMS

BY
KONSTANTINOS NTOUGIAS

DISSERTATION SUBMITTED 2019

AALBORG UNIVERSITY
DENMARK

Efficient Resource Allocation and Spectrum Utilisation in Licensed Shared Access Systems

AALBORG UNIVERSITY
DENMARK

Konstantinos Ntougias
Department of Electronic Systems
Fredrik Bajers Vej 7, 9220 Aalborg
Denmark

This dissertation is submitted for the degree of
Doctor of Philosophy

2019

Dissertation submitted: 12. september 2019

PhD supervisor: Associate Professor Troels Bundgaard Sørensen
Aalborg University

PhD committee: Associate Professor Carles Navarro Manchon (chairman)
Aalborg University

Professor Ana García Armada
Universidad Carlos III de Madrid

Dr. Michalis Matthaiou
Queens University of Belfast

PhD Series: Technical Faculty of IT and Design, Aalborg University

Department: Department of Electronic Systems

ISSN (online): 2446-1628
ISBN (online): 978-87-7210-502-4

Published by:
Aalborg University Press
Langagervej 2
DK – 9220 Aalborg Ø
Phone: +45 99407140
aauf@forlag.aau.dk
forlag.aau.dk

Printed in Denmark by Rosendahls, 2019

This document was typeset by the author using LaTeX2$_\epsilon$ with the *TexMaker 5.2.0* editor. The simulations were run in *MATLAB® R2018a* and in *Java* using the *Eclipse IDE*. The figures were created in *Microsoft PowerPoint* and in *MATLAB® R2018a*. The lineal icons in Fig. 1.1 and Fig. 1.5 have been made by Good Ware from www.flaticon.com and are free for use under the requirement to attribute the author / artist and refer the cite (for more information, see www.flaticon.com).

To my parents.

Abstract

The vision of the fifth generation (5G) standard of cellular mobile radio communications entails the provision of a multitude of "data-hungry" services. At the same time, though, the highly congested sub-6 GHz spectrum is expected to play a key role in the landscape of future cellular networks, as a means to provide the required radio coverage and mobility support. Spectrum sharing has been proposed as a countermeasure to the spectral scarcity. However, this paradigm has been met with skepticism by the operators and vendors, due to its lack of quality-of-service (QoS) provisioning. Licensed shared access (LSA) addresses this issue by enforcing orthogonal access on the shared spectrum on a non-interfering basis. Nevertheless, the enormous capacity demands of 5G networks call for more efficient sharing of the spectrum.

To this end, the combination of multi-cell multi-user multiple-input multiple-output (MU-MIMO) technologies – i.e., coordinated multi-point (CoMP) and massive MIMO (mMIMO) – with underlay spectrum sharing promises substantial spectral efficiency (SE) gains and QoS guarantees to the end users, thanks to the advanced resource allocation and interference management features of these technologies. Therefore, this paradigm could complement LSA into a next-generation LSA framework, to extend the usable spectrum. This concept, though, has been largely overlooked in the literature, in our utmost surprise. Moreover, the few relevant works in underlay spectrum sharing based on CoMP neglect the QoS requirements of the end users or the user selection procedure. Also, the majority of these studies does not consider the application of standard linear precoding schemes, which are well-known and robust precoding solutions, as a means to accelerate the adoption of this spectrum sharing paradigm by commercial deployments. Furthermore, the performance of CoMP is limited by the number of base station (BS) antennas. In addition, the joint transmission (JT) variant of CoMP is rarely utilized in practice, due to the heavy burden that it imposes on the mobile transport network (fronthaul / backhaul) in terms of capacity and latency requirements.

There is also a recent interest in the sharing of millimeter-wave (mmWave) spectrum, due to the high distance-dependent path loss and probability of blo-

ckage in such high frequencies that facilitates interference management. Efficient hybrid analog-digital precoding techniques for mMIMO setups are required in this case.

In this dissertation, we aspire to fill these gaps in the literature. The contributions of this work are summarized as follows:

- QoS-aware and QoS-agnostic coordinated power allocation (PA) schemes are derived for sum-rate (SR) maximization, under the assumption that standard linear precoding schemes, such as zero-forcing (ZF) and regularized ZF precoding, are utilized. The QoS requirements refer to the minimum rate constraints (MRC) of the users. These techniques take the form of multi-level water-filling (WF) power allocation. Simple suboptimal coordinated PA strategies are proposed too.
- Projected ZF precoding for various primary system (PS) setups is derived.
- Low-complexity heuristic coordinated user selection (CS) methods based on reduction of the search space or on the cross-correlation of the user channels with the inter-system interference channels are presented. Greedy implementations are also described.
- Cache-aided CoMP-JT techniques are developed – in particular, a coordinated caching strategy that creates JT opportunities and two caching schemes that increase the cache hit rate in comparison to the "de-facto standard" least recently used (LRU) scheme, yet maintain its $\mathcal{O}(1)$ complexity.
- Coordinated ZF precoding for load-controlled antenna arrays is derived, to improve the performance for a given number of radio frequency (RF) units. Also, a beam selection and precoding (BSP) approach is proposed, as a workaround to the difficulties posed by load computation for precoding.
- Coordinated symbol-level precoding is presented. This technique improves the performance in the low signal-to-noise-ratio (SNR) regime.
- Finally, hybrid processing via stochastic approximation with Gaussian smoothing (HPSAGS) is derived, which provides good performance with low computational complexity. This method is studied in millimeter-wave mMIMO setups.
- These techniques are evaluated via an extensive set of numerical simulations. In system-level simulations of CoMP setups a proposed dynamic cell clustering scheme is utilized. The 3GPP non-line-of-sight (NLOS) macro-cellular model is considered.

Our study shows that underlay spectrum sharing with QoS guarantees is possible over a wide range of interference power thresholds (IPT) and receive SNRs in

both sub-6 GHz and mmWave spectrum. The supported data rates (QoS) depend on the aforementioned factors, with higher data rates being possible at the high SNR regime for moderate or relaxed IPTs. We expect that this work will motivate further research on this topic.

Dansk Resumé

Visionen om femte generation (5G) cellulær mobilkommunikation baserer sig på et stort udbud af datakrævende services. Samtidig er der en forventning om, at det meget trængte frekvensbånd under 6 GHz stadig spiller en central rolle i fremtidige mobilkommunikationsnetværk, for at kunne tilbyde den ønskede dækning og mobilitets support.

Spektrumdeling har været foreslået som et modtræk til et trængt frekvensspektrum. Det er dog blevet mødt med kritik og skepsis af operatører og leverandører, grundet manglen på Quality-of-Service (QoS) kontrol. Licensed Shared Access (LSA) addresserer denne problematik ved at indføre ikke-interfererende ortogonal adgang til det delte spektrum. I betragtning af den krævede kapacitet i 5G netværk, kræves der dog mere effektiv deling af spektrum end hvad der hidtil har været muligt.

Et lovende koncept i denne sammenhæng er multi-celle Multi-User Multiple-Input-Multiple-Output (MU-MIMO) teknologier - eksempelvis Coordinated Multi-Point (CoMP) og massive MIMO (mMIMO) - i kombinationen med overlappende spektrumdeling, hvilket muliggør reelle gevinster i form af spektraleffektivitet (SE) og QoS garantier til slutbrugerne, sidstnævnte muliggjort af avanceret resourceallokering og interferenshåndtering. Kombinationen kan komplementere LSA i et næste-generations LSA koncept, og derved sikre bedre udnyttelse af det tilgængelige frekvensspektrum.

Konceptet er dog, overraskende, stort set overset i den tekniske litteratur. De få studier der er tilgængelige omkring overlappende spektrumdeling baseret på CoMP, tilsidesætter enten slutbrugernes QoS krav eller udvælgelsen af overlappende brugere. Ydermere betragter flertallet af sådanne studier ikke anvendelsen af standard lineær prekodningsteknikker, der ellers er velkendte og robuste teknikker, til at sikre en hurtig accept og implementering af spektrumdeling i kommercielle netværk. Årsagerne kan ligge i at gevinsten ved CoMP teknikken er begrænset af antallet af basisstation (BS) antenner, og at Joint Tranmission (JT) varianten af CoMP sjældent anvendes i praksis pga. den anseelige belastning af transportnetværket (fronthaul / backhaul i mobilnetværket) der fremkommer af kravene til kapacitet og latenstid i netværket.

Der er også en nylig interesse i deling af millimeter bølge (mmWave) frekvensspektret, grundet den høje afstandsafhængighed af udbredelsestabet og sandsynligheden for blokering af signalet, som begge gavner i forbindelse med håndtering af interferens. Effektive hybride analog-digital prekodningsteknikker for mMIMO er krævet i dette tilfælde.

Målet i denne afhandling er at bidrage med ovenstående manglende studier, og demed fylde dette hul i den tilgængelige litteratur. Bidragene i afhandlingen er følgende:

- Udledning af QoS-aware og QoS-agnostic power allokerings (PA) teknikker for sum-rate (SR) maksimering, under antagelse om at standard lineære prekodningsteknikker, såsom zero-forcing (ZF) og regularized ZF prekodning, anvendes. QoS kravene refererer her til krav om minimum datarate for slutbrugerne. Teknikkerne resulterer i multi-level water-filling (WF) power allokeringsteknikker, samt strategier for simpel suboptimal koordineret PA.

- Udledning af Projected ZF prekodning for forskellige eksempler af primære systemer.

- Præsentation af heuristisk udledte lav-kompleksitetsmetoder til koordineret udvælgelse af overlappende brugere, baseret på reducering af søgerummet eller på krydskorrelationen mellem brugernes kanaler og inter-system interferenskanalerne. Grådige implementeringer er også beskrevet.

- Udvikling af cache-aided CoMP-JT teknikker, herunder specifikt en koordineret caching strategi der tilvejebringer JT muligheder, og to caching metoder der forøger hitraten i sammenligning med state-of-the-art least recently used (LRU) metoden, men som bibeholder $\mathcal{O}(1)$ kompleksitet.

- Udledning af koordineret ZF prekodning for antenna arrays med kontrolleret passiv belastning, med henblik på at forbedre performance givet et begrænset antal radiofrekvens (RF) enheder; der foreslåes også en beam selection og prekodnings (BSP) teknik for at omgå kompleksiteten i forbindelse med beregning af den passive belastning på antenne array'et.

- En teknik til koordineret prekodning på symbolniveuau, som forbedrer performance under lave signal-til-støj-forhold (SNR)

- Udledning af processeringsmetode – hybrid processing via stochastic approximation with Gaussian smoothing (HPSAGS) – der giver good performance med lav beregningsmæssig kompleksitet. Denne metode studeres i mmWave-mMIMO-opsætninger.

- Disse teknikker evalueres via et omfattende sæt numeriske simuleringer. I simuleringer på systemniveau af CoMP-opsætninger bruges et foreslået dy-

namisk celleklyngsskema. Den makrocellulære model 3GPP ikke-synsvinkel (NLOS) overvejes.

Studiet viser at overlappende spektrumdeling med QoS garantier er muligt for et stort spænd af værdier for interferenstærskel, eller interference power threshold (IPT), og modtaget SNR, gældende for spektrum både under 6 GHz og i mmWave. De supporterede datarater (QoS) afhænger af de førnævnte faktorer, hvor højere datarater er mulige for de høje SNR og moderate eller høje IPTs. Forventningen er, at afhandlingens resultater vil motivere til yderligere forskning omkring emnet.

Acknowledgements

I would like to express my gratitude to my supervisors, Associate Professor Troels B. Sorensen, Professor Klaus I. Pedersen, and Adjunct Professor Constantinos B. Papadias, for their continuous guidance and encouragement as well as for their invaluable advices throughout these years.

I am also indebted to all my colleagues and co-authors for our great cooperation. Special thanks go to Dr. Dimitrios Ntaikos, who answered any question I had (and some that I didn't have) about antenna arrays in general and load-controlled parasitic antenna arrays in particular and simulated numerous antenna array designs; Dr. Georgios K. Papageorgiou, for introducing to me the world of optimization; and Dr. Gerhard Hasslinger, who gave me the opportunity to study the concept of caching.

Finally, I would like to thank my parents, for their unconditional love; George and Maria, for their friendship; and Christina, for her continuous support.

Thesis Details

Thesis Title:	Efficient Resource Allocation and Spectrum Utilisation for Licensed Shared Access Systems
Ph.D. Student:	Konstantinos Ntougias
Supervisor Team:	Associate Professor Troels B. Sorensen, Department of Electronic Systems, Aalborg University, Denmark
	Professor Klaus I. Pedersen, Department of Electronic Systems, Aalborg University, Denmark
	Adjunct Professor Constatinos B. Papadias, Department of Electronic Systems, Aalborg University, Denmark
PhD Committee:	Associate Professor Carles N. Manchon (chairman), Department of Electronic Systems, Aalborg University, Denmark
	Professor Ana G. Armada, Department of Signal Theory and Communications, Universidad Carlos III de Madrid, Madrid, Spain
	Professor Michalis Matthaiou, Institute of Electronics, Communications and Information Technology, Queen's University Belfast, UK

This Thesis is based on the following scientific work:
Conference Papers

1. K. Ntougias, D. Ntaikos, C. B. Papadias, "Robust low-complexity arbitrary user- and symbol-level multi-cell precoding with single-fed load-controlled parasitic antenna arrays", *Int. Conf. Telecommun. (ICT)*, Thessaloniki, Greece, May 16-18, 2016.

2. K. Ntougias, D. Ntaikos, C. B. Papadias, "Coordinated MIMO with single-fed load-controlled parasitic antenna arrays," *IEEE Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Edinburgh, UK, July 3-6, 2016.

3. K. Ntougias, D. Ntaikos, B. Gizas, G. K. Papageorgiou, C. B. Papadias, "Large load-controlled multiple-active multiple-passive antenna arrays: Transmit beam-forming and multi-user precoding," *Eur. Signal Process. Conf. (EUSIPCO)*, Kos, Greece, 28 Aug.-2 Sept., 2017.

4. K. Ntougias, D. Ntaikos, C. B. Papadias, "Single-and multiple-RF load controlled parasitic antenna arrays operating at Cm-wave frequencies: Design and applications for 5G wireless access/backhaul," *FITCE Congress*, Madrid, Spain, Sept. 14-15, 2017.

5. K. Ntougias, D. Ntaikos, C. B. Papadias, G. K. Papageorgiou, "Simple cooperative transmission schemes for underlay spectrum sharing using symbol-level precoding and load-controlled arrays," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 7854-7858, 2019.

6. K. Ntougias, D. Ntaikos, C. B. Papadias, G. K. Papageorgiou, "Coordinated hybrid precoding and QoS-aware power allocation for underlay spectrum sharing with load-controlled antenna arrays," *IEEE Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Cannes, France, July 2-5, 2019.

7. K. Ntougias, C. B. Papadias, G. K. Papageorgiou, G. Hasslinger, "Spectral Coexistence of 5G Networks and Satellite Communication Systems Enabled by Coordinated Caching and QoS-Aware Resource Allocation," *Eur. Signal Process. Conf. (EUSIPCO)*, A Coruna, Spain, September 2-6, 2019 (to appear).

Journal Articles

1. G. Hasslinger, K. Ntougias, F. Hasslinger, O. Hohlfeld, "Performance evaluation for new web caching strategies combining LRU with score based object selection," *Computer Netw.*, vol. 125, Oct. 2017, pp. 172–186.

2. K. Ntougias, C. B. Papadias, G. K. Papageorgiou, G. Hasslinger, T. B. Sorensen, "Coordinated caching and QoS-aware resource allocation for spectrum sharing," *Springer Wireless Commun. Netw.* (to appear).

3. G. K. Papageorgiou, M. Sellathurai, K. Ntougias, C. B. Papadias, "A stochastic optimization approach to hybrid processing in massive MIMO systems," *IEEE Wireless Commun. Letters* (submitted, awaiting review).

4. K. Ntougias, G. K. Papageorgiou, C. B. Papadias, T. Sorensen, M. Sellathurai, "Low-complexity coordinated resource allocation for QoS-constrained SR maximization in Underlay Specrum Sharing Setups," *IEEE Trans. Wireless Commun.* (submitted, awaiting review).

Book Chapters

1. K. Ntougias, D. Ntaikos, C. B. Papadias, "Channel-dependent precoding for multiuser access with load-controlled parasitic antenna arrays," in *New directions in wireless communications systems: From mobile to 5G*, A. G. Kanatas, K. S. Nikita, P. Mathiopoulos (Eds.), CRC Press, 2017, ch. 8, pp. 279-314.

2. K. Ntougias, D. Ntaikos, G. K. Papageorgiou, C. B. Papadias, "Interference avoidance and mitigation techniques for hybrid satellite-terrestrial networks," in *Satellite Communications in the 5G Era*, S. K. Sharma. S. Chatzinotas, P. D. Arapoglou (Eds.), IET, 2018, ch. 16, pp. 459-490.

3. C. B. Papadias, K. Ntougias, G. K. Papageorgiou, "The role of antenna arrays in spectrum sharing," in *Spectrum Sharing: The Next Frontier in Wireless Networks*, C. B. Papadias, T. Ratnarajah, D. T. M. Slock (Eds.), Wiley, 2019, ch. 12 (submitted, awaiting review).

This Thesis has been submitted for assessment in partial fulfillment of the PhD degree. The co-author statements have been made available to both the assessment committee and Doctoral School of Aalborg University.

Contents

List of Figures

List of Tables

Abbreviations

3GPP	3rd Generation Partnership Project
5G	5th Generation
AAE	Active AE
ACI	Adjacent Channel Interference
ADC	Analog to Digital Converter
AE	Antenna Element
AM	Amplitude Modulation
AoA	Angle-of-Arrival
AoD	Angle-of-Departure
AR	Augmented Reality
ASA	Authorized Shared Access
AWGN	Additive White Gaussian Noise
BC	Broadcast Channel
BER	Bit Error Rate
BF	Beamforming
BLER	Block Error Rate
BPSK	Binary Phase Shift Keying
BS	Base Station
BSP	Beam Selection and Precoding
C3RE	Coordinated Content Caching w/ Redundancy Enhancement
CBF	Coordinated BF
CBRS	Citizens Broadband Radio Service
CCI	Co-Channel Interference
CDSA	Control Plane / Data Plane Separation Architecture
CEPT	Eur. Conf. of Postal & Telecommun. Administrations
CI	Constructive Interference
C-ICEPA	Coordinated ICEPA
C-ICPA	Coordinated ICPA
CIR	Channel Impulse Response
C-IUEPA	Coordinated IUEPA
C-IUPA	Coordinated IUPA

CLT	Central Limit Theorem
CN	Core Network
CO	Central Office
CoMP	Coordinated Multi-Point
CP	Coordinated Precoding
CPA	Coordinated PA
CQA-ICPA	Coordinated QoS-aware ICPA
CQA-IUPA	Coordinated QoS-aware IUPA
CR	Cognitive Radio
CS	Coordinated Scheduling
CSI	Channel State Information
CSIR	CSI at RX
CSIT	CSI at TX
CTF	Channel Transfer Function
C-RZF	Coordinated RZF
C-ZF	Coordinated ZF
DAC	Digital to Analog Converter
DCC	Dynamic Cell Clustering
DCS	Dynamic Cell Selection
DoF	Degrees of Freedom
DL	Downlink
DPC	Dirty Paper Coding
EB	Exabyte
eLSA	Evolved LSA
EM	Electromagnetic
EPA	Equal PA
ETSI	European Telecommunications Standards Institute
FCC	Federal Communications Commission
FD	Full Duplex
FDD	Frequency Division Duplex
FDMA	Frequency Division Multiple Access
FEC	Forward Error Correction
FFR	Fractional Frequency Reuse
FIS	Forward Inter-System
FM	Frequency Modulation
FSPL	Free Space Path Loss
GB	Gigabyte
Gbps	Gigabits per second
GSM	Global System for Mobile Communications
GHR	Global HR
GHz	Gigahertz
HD	Half Duplex
HetNet	Heterogeneous Network
HPSAGS	Hyprid Precoding via Stochastic Approx. w/ Gaussian Smoothing

HR	Hit Rate
Hz	Hertz
IBC	Interference BC
IBFD	In-Band Full Duplex
IC	Interference Channel
ICEPA	Interference Constrained EPA
ICI	Inter-Cell Interference
ICPA	Interference Constrained PA
i.i.d.	Independent and identically distributed
IO	Interacting Object
IRM	Independent Reference Model
IPC	Interference Power Constraint
IPT	Interference Power Threshold
ISD	Inter-Site Distance
ISI	Inter-Symbol Interference
ISM	Industrial, Scientific, and Medical
ITU	International Telecommunication Union
IUEPA	Interference Unconstrained EPA
IUPA	Interference Unconstrained PA
JT	Joint Transmission
kHz	Kilohertz
KKT	Karush-Kuhn-Tucker
KPI	Key Performance Indicator
LC-PAA	Load-Controlled Parasitic Antenna Array
LFU	Least Frequently Used
LHR	Local HR
LOS	Line-of-Sight
LRU	Least Recently Used
LSA	Licensed Shared Access
LTE	Long Term Evolution
LTE-A	LTE-Advanced
LTV	Linear Time-Variant
MAMP	Multiple Active Multiple Passive
MB	Megabyte
MBB	Mobile Broadband
MBH	Mobile Backhaul
Mbps	Megabits per second
MFH	Mobile Fronthaul
MHz	Megahertz
MIMO	Multiple Input Multiple Output
mMIMO	Massive MIMO
MMSE	Minimum Mean Square Error
MNO	Mobile Network Operator
MPC	Multi-Path Component
MRC	Minimum Rate Constraint
MRT	Maximum Ratio Transmission

MS	Mobile Station
MUI	Multi-User Interference
MU-MIMO	Multi-User MIMO
MVND	Matrix Variate Normal Distribution
MVNO	Mobile Virtual Network Operator
mmWave	Millimeter Wave
NC	Network-Centric
NLOS	Non-LOS
NRA	National Regulatory Administration
OBFD	Out-of-Band Full Duplex
OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal FDMA
OOC	Out-of-Cluster
OSA	Opportunistic Spectrum Access
PA	Power Allocation
PAE	Passive AE
p.d.f.	Probability Density Function
PLE	Path Loss Exponent
PMSE	Program Making and Special Equipment
PS	Primary System
PSD	Power Spectral Density
PU	Primary User
P-ZF	Projected ZF
QoS	Quality of Service
RA	Resource Allocation
RAN	Radio Access Network
RF	Radio Frequency
RHS	Right Hand Side
RIS	Reverse Inter-System
RRU	Remote Radio Unit
RSS	Reduced Search Space
RX	Receiver
RZF	Regularized ZF
SAMP	Single Active Multiple Passive
SAS	Secondary Access System
SDMA	Space Division Multiple Access
SE	Spectral Efficiency
SER	Symbol Error Rate
SG-C	Score-Gated Clock
SGD	Stochastic Gradient Descent
SG-LRU	Score-Gated LRU
SI	Self-Interference
SINR	Signal-to-Interference-plus-Noise-Ratio
SIR	Signal-to-Interference-Ratio
SISO	Single Input Single Output
SM	Spatial Multiplexing

SMS	Short Message Service
SNR	Signal-to-Noise-Ratio
SPC	Sum Power Constraint
SR	Sum Rate
SS	Secondary System
SSPOMP	Spatially Sparse Precoding w/ Orthogonal Matching Pursuit
SU	Secondary User
SVD	Singular Value Decomposition
TB	Terabyte
TCO	Total Cost of Ownership
TDD	Time Division Duplex
TDMA	Time Division Multiple Access
THR	Total HR
THz	Terahertz
TRX	Transceiver
TV	Television
TVWS	TV White Spaces
TX	Transmitter
UDN	Ultra Dense Network
UHD	Ultra High Definition
UC	User-Centric
UK	United Kingdom
UL	Uplink
ULA	Uniform Linear Array
UN	United Nations
U.S.A.	United States of America
vBBU	Virtual Baseband Unit
VR	Virtual Reality
Wi-Fi	Wireless Fidelity
WF	Water Filling
WLFU	Window LFU
WLFU-NE	WLFU w/ Neighbor Exchange
WRC	World Radiocommunication Conference
w.r.t.	with respect to
WSR	Weighted SR
WSSUS	Wide Sense Stationary Uncorrelated Scattering
ZF	Zero Forcing

Symbols

$\mathbb{C}$	Set of complex numbers
$\mathbb{R}$	Set of real numbers
$\mathbb{R}_+$	Set of non-negative reals
$a \in \mathbb{C}$	Complex-valued scalar
$\mathrm{Re}\{a\}$	Real part of complex-valued scalar a
$\mathbf{a} \in \mathbb{C}^n$	n-dimensional complex-valued vector
$\lvert a \rvert$	Magnitude (absolute value) of complex (real) scalar a
$\mathbf{A} \in \mathbb{C}^{n \times m}$	$n \times m$ matrix $\mathbf{A}$ with complex-valued entries
$(\mathbf{a})_i = a_i$	i-th element of $\mathbf{a}$
$(\mathbf{A})_{ij} = a_{ij}$	(i,j)-th entry of $\mathbf{A}$
$(\mathbf{A})_{i*}$	i-th row of $\mathbf{A}$
$(\mathbf{A})_{*j}$	j-th column of $\mathbf{A}$
$\lVert \mathbf{a} \rVert$	Euclidean norm of $\mathbf{a}$
$\lVert \mathbf{A} \rVert$	Euclidean norm (i.e., 2-norm) of $\mathbf{A}$
$\lVert \mathbf{A} \rVert_F$	Frobenius norm of $\mathbf{A}$
$\mathbf{A} = \mathrm{diag}\,(a_1, \ldots, a_n)$	Diagonal matrix $\mathbf{A}$ with on-diagonal entries $a_{ii} = a_i$
$\mathbf{A}^T$	Transpose of $\mathbf{A}$
$\mathbf{A}^*$	Conjugate of $\mathbf{A}$
$\mathbf{A}^\dagger$	Hermitian (conjugate transpose) of $\mathbf{A}$
$\mathbf{A}^{-1}$	Inverse of $\mathbf{A}$
$\mathbf{A}^\#$	Moore-Penrose pseudo-inverse of $\mathbf{A}$
$\mathrm{vec}\,(\mathbf{A})$	Vectorization of $\mathbf{A}$
$\mathbf{I}_n$	$n \times n$ identity matrix
$\mathbf{O}_{n \times m}$	$n \times m$ zero matrix
$\mathbf{0}_n$	n-dimensional null vector
$\mathbf{A} \otimes \mathbf{B}$	Kronecker product of $\mathbf{A}$ and $\mathbf{B}$
$\mathbf{A} \odot \mathbf{B}$	Handamard (element-wise) product of $\mathbf{A}$ and $\mathbf{B}$
$\det(\mathbf{A})$	Determinant of $\mathbf{A}$
$\mathrm{tr}\,(\mathbf{A})$	Trace of $\mathbf{A}$
$a \sim \mathcal{N}\left(0, \sigma^2\right)$	Real Gaussian variable (mean, variance)
$a \sim \mathcal{CN}\left(0, \sigma^2\right)$	Complex Gaussian variable (mean, variance)

$\mathbf{a} \sim \mathcal{CN}\left(\mathbf{0}_n, \sigma^2 \mathbf{I}_n\right)$	Complex Gaussian vector (mean, variance)
$\mathbb{E}\left(\cdot\right)$	Expectation operator
$p(x\|y)$	Conditional probability of x given y
∇f	Gradient of function f
$x * y$	Convolution of x with y
$\mathcal{S} = \{S_{\min}, \ldots, S_{\max}\}$	Ordered set of integers
$\binom{n}{r}$	Combinations of r objects from a set of n objects
$\|\mathcal{S}\|$	Cardinality of set $\mathcal{S}$
$\varnothing$	Empty set
$\mathcal{A} \setminus \mathcal{B}$	The set of elements in $\mathcal{A}$ but not in $\mathcal{B}$
a^+	Maximum between 0 and a

Chapter 1

Introduction

Wireless communication technology has revolutionized communication, education, commerce, transportation, entrepreneurship, and entertainment, among other aspects of our daily lives. *Cellular mobile radio communication networks* play a central role in this reshaping of the society, by providing services to users on a move almost anywhere and anytime in a reliable manner. These include mobile telephony service, short message service (SMS), and various mobile Internet services, such as e-mail, instant messaging, online social networking, web browsing, file transfer, video and audio streaming, and video telephony, just to name a few.

The performance of cellular networks is quantified via a set of metrics or *key performance indicators (KPI)*, such as the peak data rate of the users and the end-to-end latency. A new generation of *cellular mobile radio communication standards* is commonly introduced every ten years or so [1], as a consequence of: (i) technological advances; (ii) major breakthroughs in communication theory, signal processing, and other relevant fields of engineering and science; and (iii) market needs. Each generation typically demonstrates improved performance in comparison to its predecessor and provides new innovative services as well as enhancements of known services.

Nowadays, we live in the dawn of the *5th Generation* (*5G*) era. Widespread deployment of commercial 5G networks is anticipated to take place worldwide around 2020. 5G envisions the support of 100–1000 times higher *capacity* than current 4G networks [1]. More specifically, 5G targets the provision of up to 100 times higher area traffic capacity (10 Mbps/m^2), 20 times higher peak data rate (20 Gbps), and 10 times higher minimum guranteed data rate (100 Mbps) [1, 2]. These goals are coupled with specific scenarios, as shown in Fig. 1.1.

These capabilities will give rise to a multitude of "data-hungry" services, such as ultra-high-definition (UHD) video streaming, virtual reality (VR), augmented reality (AR), and 3D video [2]. Therefore, the exponential growth of the mobile data traffic that has been reported over the last decade [3] (see Fig. 1.2), as a re-

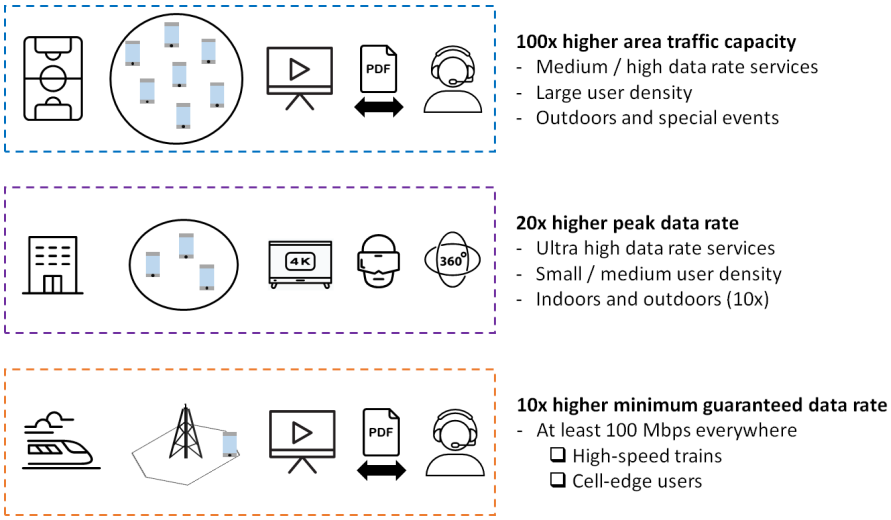


Figure 1.1 5G performance targets. [Icons: Good Ware, www.flaticon.com.]

sult of the technological evolution and the never-ending demand for higher data rates, is expected to continue in the foreseeable future. For instance, ITU forecasts that the global mobile data traffic will grow 10–100 times from 2020 to 2030, as illustrated in Fig. 1.3 [4].

The realization of the 5G vision is a highly challenging task. First of all, the wireless channel degrades the quality of communication and limits the performance of the cellular network: it attenuates and distorts the transmitted signal, adds interference by other radio signals that are transmitted concurrently with the desired signal, and contaminates the received signal with thermal background noise [5]. Additionally, the main communication resources are limited. In particular, the frequency bands that are suitable for long-range communication are shared among several applications and corresponding systems (e.g., cellular wide area networks, wireless local area networks, TV and radio broadcasting systems, satellite communication systems, etc.) [5]. Perhaps more importantly, each operator makes exclusive use of its slice of spectral resources [5]. Also, the transmission power is subject to hardware limitations, regulatory restrictions, interference constraints, and cost considerations [5]. These facts should be taken into account in the design of the next-generation cellular networks, which will have to cope with an explosion of the mobile data traffic.

Hence, the following question naturally arises:

How can we achieve the ambitious 5G capacity goal and meet the stringent requirements of the envisaged services, given the hostility of the wireless channel and the limited availability of resources?

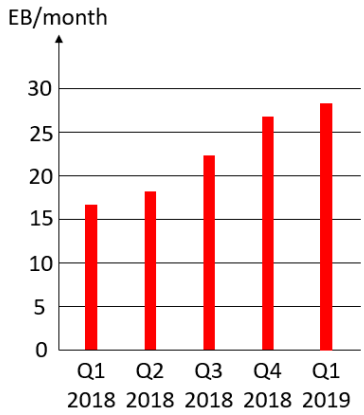


Figure 1.2 Total (uplink and downlink) global mobile data traffic per quarter in EB/month over the last year. (Note: The values in the plot might differ slightly from the corresponding values reported by Ericsson in [3].)

The main objective of this chapter is to provide a deeper understanding of this question, take a glance at its answer, and highlight the focal point of this dissertation. To this end, the chapter begins with a short review of some fundamental concepts. The purpose of these introductory sections is to present the basic principles of cellular networks and shed light on the limitations of wireless communication and the impairments of the mobile radio channel. The chapter continues with an overview of the proposed solutions to the problem stated in the above question. The challenges, misconceptions, and opportunities associated with the implementation of these solutions are described in the subsequent section. Then, the concept of resource allocation, which is of central importance in this work, is described. Next, the motivation behind our study is stated, together with our goals. The chapter concludes with the list of our contributions and the description of the structure of this monograph.

1.1 Cellular Mobile Radio Communication Systems

In wireless communication networks, devices exchange data in the form of *electromagnetic* (*EM*) *waves* that propagate through the environment. Thus, such a network can be viewed in general as a collection of *radio links*, where each one of them is comprised by a *transmitter* (*TX*), a *receiver* (*RX*), and a *communication channel* (physical medium) that links these nodes together [6].

A cellular mobile radio communication network refers to a system of interconnected nodes (network) that provides communication services to mobile devices via the transmission of radio signals. The service area is divided into a number

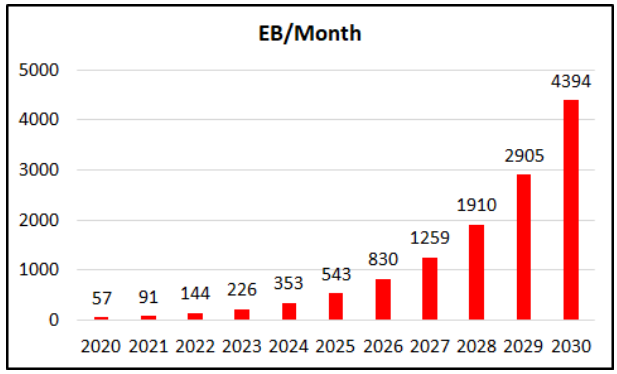


Figure 1.3 Estimations of the global mobile traffic from 2020 to 2030 (machine-to-machine traffic excluded) [4].

of smaller radio coverage segments called cells—hence the name cellular for this type of networks.

In highly simplified terms, a cellular network consists of a set of *base stations* (*BS*) that are distributed across the geographic region of interest[1]. These nodes are essentially *transceivers* (*TRX*). Each BS provides radio coverage over a *cell* and enables a number of *mobile stations* (*MS*) that are associated with it (e.g., smartphones) to access the network. The cells are classified into *macro-cells* and various types of *small cells* (micro-cells, pico-cells, femto-cells) according to their radius, which ranges from a few tens of meters to a few tens of kilometers [7]. This structure enables the network to meet a set of diverse radio coverage and throughput requirements.

Cellular networks support bi-directional communication [8]: on the *downlink* (*DL*), a BS transmits signals to its assigned users, whereas on the *uplink* (*UL*) it receives signals from its respective users. In other words, on the cellular DL the BSs and the MSs act as TXs and RXs respectively, while on the cellular UL these roles are reversed, as depicted in Fig. 1.4. In this work, we focus on the cellular DL.

1.2 Wireless Communucication Resources

1.2.1 Radio Spectrum

As *radio spectrum* is defined a subset of the EM spectrum that lies in the 3 kHz–300 GHz frequency range [9]. The radio spectrum supports a wide variety of services, from navigation to radio astronomy and from land mobile communications to

[1]In practice, there exist also other network elements and registries / databases that facilitate user authentication, mobility management, interconnection with external networks, etc.

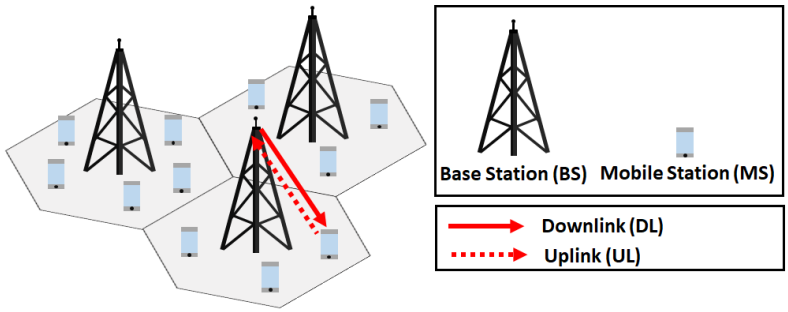


Figure 1.4 Simplified representation of a cellular network.

radars, as shown in Fig. 1.5. It constitutes the means that makes possible wireless communication, in a similar way that copper cables and optical fibers enable wireline communication, and represents a valuable resource for establishing economic growth and social development [10].

Historically, the radio spectrum has been highly regulated to prevent the occurrence of harmful interference (i.e., to maintain the interference levels below a threshold) [9, 10]. Other goals of *spectrum regulation* include the optimization of spectrum usage; the facilitation of technological standardization and spectrum harmonization, which give rise to economies of scale; the accelerated introduction of new technologies; the promotion of the public interest; and the maximization of the social benefit in general [9, 10].

Traditionally, *spectrum management* is comprised by two stages. First, the spectrum is divided into frequency bands and the purpose of these bands is determined, i.e., each band is associated with one or more services. This process is called *spectrum allocation* and is performed at international level by the International Telecommunication Union (ITU) [10]. ITU, which is a specialized agency of the United Nations (UN), reviews and, if necessary, revises the frequency allocation at the World Radiocommunication Conference (WRC) that is held every three to four years [11]. Note that the ITU allocates frequency bands to services (e.g., mobile terrestrial service), not to applications (e.g., cellular networks) neither to technologies (e.g., LTE) [11]. Notice also that it issues guidelines for the use of the spectrum in different regions and countries [12].

Spectrum allocation is followed by the assignment of specific frequency blocks comprised by one or more channels to the different operators. The channel width is application-specific, e.g., AM radio uses 10 kHz wide channels, whereas GSM divides a block of 25 MHz into 125 channels with bandwidth of 200 kHz. Also, we should note that appropriate guard bands are defined during the spectrum assignment process to maintain the interference between different operators at an acceptable level [9]. *Spectrum assignment* is a responsibility of the *national regulatory administrations* (*NRA*), such as the Federal Communications Commission (FCC) in

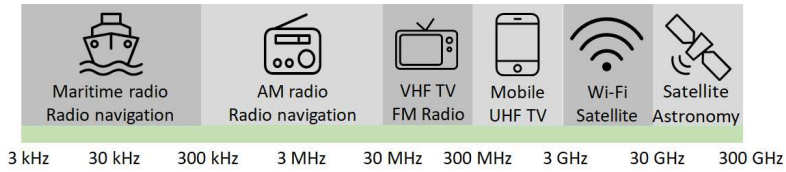


Figure 1.5 The radio spectrum supports a variety of services. [Icons: Good Ware, www.flaticon.com.]

the U.S.A. and the Office of Communication (Ofcom) in the UK [12]. Nevertheless, this process is tightly coupled with the guidelines given by the ITU and corresponding regional administrations, such as the European Conference of Postal and Telecommunications Administrations (CEPT) [12]. In any case, spectrum assignment is consistent with spectrum allocation (e.g., TV stations are assigned frequencies that lie within the bands allocated to broadcasting services) [11]. Often, similar frequency ranges are assigned to similar applications by the respective NRAs—e.g., FM radio broadcasting stations operate in the frequency range 87.5–108 MHz, whereas the frequency ranges 800–1000 MHz and 1800–2100 MHz are typically utilized by cellular communication networks [12]. We should mention that spectrum assignment is accompanied by the definition of rules and conditions regarding spectrum usage.

Initially, spectrum assignment was performed in an administrative manner, i.e., the governments assigned licenses to particular users. This form of *spectrum authorization* was commonly based on comparative evaluation ("beauty contests"), where the applicants had to guarantee compliance with certain requirements (e.g., radio coverage, *quality-of-service* (*QoS*), etc.) to obtain a license, or was taking place on a first-come first-served basis [10, 12]. Today, *administrative licensing* has been largely replaced by *spectrum auctions*, where the applicants bid for licenses [10]. This strategy leads to more efficient use of the spectrum and provides a significant revenue to the governments from its utilization, over which they have sovereign rights in the respective countries.

The spectrum licenses typically grant their owners the right to use exclusively one or more frequency blocks over large geographical areas (usually country-wide) and long time periods (commonly for many decades), in order to provide a radio communication service under specified spectrum usage rules and conditions that govern the corresponding application. This *licensed spectrum access* model ensures interference-free operation and enables the provision of predictable QoS. Hence, it justifies long-term investments on network infrastructure [13].

On the other hand, the demand for spectrum has grown significantly over the last decades, as illustrated by the plethora of applications in today's radio communication landscape and the recent advent of *mobile broadband* (*MBB*) services. Under this emerged reality, the radio spectrum has become a scarce (and, conse-

quently, expensive) resource. This so-called *spectrum crunch* issue is mainly noticed in the highly congested sub-6 GHz segment, which has been traditionally utilized for long-range communication purposes due to its favorable propagation characteristics. The shortage in spectral resources is attributed to the rigid nature of the licensed spectrum access paradigm [9, 10]. The inefficiency of this spectrum management model is further highlighted by the low utilization of the assigned frequency blocks in the space, time, and frequency domains (e.g., see the study carried out by Analysis Mason for the European Commission in 2013 [14]).

In order to deal with this somewhat artificial spectrum scarcity, the European Conference of Postal and Telecommunications Administrations (CEPT) in Europe and the FCC in the U.S.A. promote the shared use of the spectrum on a non-interfering basis [13]. This typically requires the utilization of a geolocation database and, possibly, the application of spectrum sensing to detect the activity of incumbents and infer the radio propagation conditions. Other approaches besides *spectrum sharing* include the re-purposing or *re-farming* of legacy spectrum, as in the digital dividend use case where the transition of TV broadcasting from analog to digital technology freed up spectrum in favor of MBB and other services [15], as well as the utilization of higher frequencies, where substantial amounts of unexploited bandwidth can be found [16]. One should mind, though, that at extremely high frequencies the transmitted radio waves are severely attenuated with the propagation distance and get easily blocked by objects. Thus, such frequencies can be used only for short-range communication. Spectrum re-farming, on the other hand, is a long and complex procedure that can "unlock" the access to a limited amount of spectral resources.

A small portion of the radio spectrum has been reserved for *unlicensed access* or "*license-exempt*" use under given restrictions (e.g., on the transmission power levels and the geographic areas) whose goal is to limit or avoid interference [10]. Although the use of unlicensed spectrum for short-range communication in indoor environments represents a success story, as exemplified by the rise of the Wi-Fi standard which operates in the unlicensed 2.45 GHz and 5.8 GHz ISM bands [10], this spectrum usage paradigm has not been adopted for wide-area coverage and service provision by *mobile network operators* (*MNO*). The reason is that one cannot ensure protection from harmful interference or guarantee a certain QoS level under this spectrum management model (i.e., the users can enjoy only "*best-effort*" service).

1.2.2 Transmission Power

The transmission power is subject to a number of constraints [5, 12]:

- **Hardware limitations:** The *radio frequency* (*RF*) power amplifier accounts for a considerable fraction of the power consumption in the transmitter. To this end, the BSs make use of power amplifiers with high efficiency. These ampli-

fiers, though, are highly non-linear. In order to avoid the occurrence of non-linear distortion, the BSs restrict their operation within the linear segment of the dynamic range. This fact limits the maximum transmission power.

- **Regulatory restrictions:** Typically, the maximum transmission power is limited by regulations to minimize the risk of health issues related with radio-wave exposure, such as extensive heating, as well as to avoid harmful interference with other systems.

- **Interference constraints:** The MNOs set the maximum transmission power of the BSs such that the interference levels at the network are acceptable.

- **Cost considerations:** The maximum transmission power is determined also by the cost for running the network. The MNOs commonly try to minimize the operational expenditures as much as possible, without affecting though the system performance and the QoS.

1.3 Radio Propagation Mechanisms

As *wireless channel* is defined the physical medium through which the transmitted radio signals propagate to reach their destinations. This includes the free space as well as objects along the radio propagation path(s). These objects may refer to, for instance, buildings, cars, hills, trees, and the ground in outdoor environments or walls, windows, doors, and furniture in indoor environments. The objects in the environment impact the transmission of radio signals through a number of mechanisms, which are described below [17–21] (see Fig. 1.6):

- **Transmission and Blockage:** When a propagating radio signal impinges upon an object whose size is large in comparison to the wavelength, such as a building or a wall, it may penetrate it or get blocked by it. The former phenomenon is known as *transmission* or *penetration*, while the latter one is called *blockage*. As the EM wave passes through the object, part of its energy is absorbed by the material (and possibly part of it is reflected by the output interface of the object). Blockage refers to the extreme situation of severe attenuation that results in blocking of the radio propagation. The occurrence of transmission or blockage, as well as the amount of *absorption loss* (also called *penetration loss*), depends on the size and thickness of the obstacle, the structure of its material, and the frequency of the radio signal. The absorption loss is higher for thicker objects and for radio waves with higher frequency. Transmission is important for the establishment of communication in indoor setups, where, for example, a radio wave may penetrate a thin wall to reach the receiver.

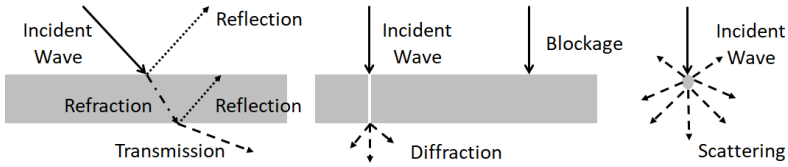


Figure 1.6 Radio propagation mechanisms.

- **Reflection:** When the propagation of a radio signal is obstructed by an object whose size is large compared to the wavelength and its surface is smooth (i.e., the dimensions of its protrusions are larger than the wavelength), such as the ground or the ceiling, the floor, and the walls in a room, then part of its energy may be reflected off this object, while the remaining part will be transmitted through the object. This phenomenon is referred to as *reflection*. The occurrence of reflection, as well as the amount of *reflection loss*, depends on the size of the object, the smoothness of its surface, the frequency and polarization of the radio signal, and the angle of incidence at the area of reflection. Generally, reflection is more intense for higher frequencies. Reflection plays an important role in both indoor and outdoor environments.

- **Refraction:** When a radio signal that is transmitted through a medium (e.g., air) enters another medium (e.g., a wall), its propagation direction changes. This is because the velocity of propagation depends on the density of the physical medium through which the radio wave propagates. This mechanism is known as *refraction*.

- **Diffraction:** When a transmitted radio signal encounters a large object (relative to its wavelength) that has sharp edges or small openings (e.g., the top of a building, a street corner, or a doorway), it may excite it, i.e., the object may act as a secondary wave source. This results in a number of weaker radio waves that propagate at different directions inside the shadow area of the obstacle. Alternatively, we could interpret this phenomenon as if part of the radiated energy bend around the object or spread out through the apertures. This propagation mechanism is called *diffraction*. Typically, diffraction results in greater power loss than reflection. On the other hand, this phenomenon is important in outdoor urban environments, where communication is commonly obstructed by large buildings. The occurrence of diffraction, as well as the amount of *diffraction loss*, depends on the frequency of the radio signal. For high frequency signals, the geometry of the object, as well as the amplitude, phase, and polarization of the incident wave at the point of diffraction play also an important role. In general, lower frequency signals penetrate deeper into the shadow region of an obstacle. This is one of the reasons why cellular networks have been utilizing sub-3 GHz spectrum.

- **Scattering:** When the propagation path consists of an object that has either a rough surface or irregular shape or when it is comprised by an object or a cluster of objects whose size is comparable to or smaller than the wavelength, then the energy of the incident radio wave is diffused into different directions. This phenomenon is referred to as *scattering*, *diffusion*, or *diffuse reflection* (as opposed to the previously described *specular reflection* caused by smooth surfaces). Examples of objects that induce scattering include street signs, lamp posts, and foliage in outdoor environments and rough walls and furniture in indoor environments. The *scattering loss* is typically larger than the reflection loss, due to the fact that in scattering the signal is spread over a wider area. Nevertheless, scattering is important in both indoor and urban outdoor environments, which are typically highly cluttered.

The obstacles in the environment that interact with the transmitted radio wave are referred to in general as *interacting objects* (*IO*) [12]. The IOs that cause reflection or scattering of the transmitted radio signal are commonly called collectively *scatterers* [22]. We should also note that diffraction is often referred to as *shadowing*, due to the fact that it enables communication with a MS that is located in the shadow area of a large obstacle [22].

Several measurements have been performed, in order to quantify the bulk attenuation that is introduced by different materials at different frequencies (e.g., for different types of partitions in indoor environments, such as plasterboard walls, concrete walls, and windows with aluminum siding [20, 23]).

Typically, the transmitted radio wave reaches the MS via multiple indirect or *non-line-of-sight* (*NLOS*) paths, thanks to its reflection, scattering, and diffraction by the IOs in the propagation environment [19]. This fact allows the establishment of communication even when there is no direct (i.e., unobstructed) or *line-of-sight* (*LOS*) path between the BS and the MS. This is very common, for example, in urban environments, where the antenna heights are commonly below the rooftops of the surrounding buildings. On the other hand though, these mechanisms attenuate the transmitted signal and give rise to multipath fading which can degrade significantly the quality of communication.

1.4 Cellular Communication Challenges

1.4.1 Noise and Interference

Noise refers to random disturbances of the received signal. We mainly consider the unavoidable thermal background noise that arises from the random thermal motion of the electrons in the electronic components of the receiving devices [24]. This type of noise is modeled as an *additive white Gaussian noise* (*AWGN*) process, i.e., additive Gaussian noise with uniform power across the entire spectrum.

The dominant limiting factor in cellular radio communications, though, is *interference*, which refers to the disruption of communication by unintended signals that are received simultaneously with the desired signal. Interference is attributed to the sharing of the communication medium among multiple transmissions [11].

Depending on whether the sources of interference are internal or external w.r.t. the cellular network, we classify interference into *intra-system interference* and *inter-system interference*, respectively.

When the interfering signals have the same frequency with the desired signal, then we refer to *co-channel interference* (*CCI*), while when they have adjacent frequencies, we refer to *adjacent channel interference* (*ACI*) [23]. The latter type of interference is caused usually by the non-ideal response of the receive filter, which results in energy leakage from such unintended signals [20].

Letting a device to transmit and receive simultaneously on the same frequency or a BS to communicate with a group of users on a single time-frequency resource or two nodes at neighboring cells to transmit in parallel over the same frequency band results in more efficient utilization of the scarce spectral resources. On the other hand, though, these strategies give rise to CCI. These types of CCI are called *UL/DL self-interference* (*SI*), *intra-cell CCI* or *multi-user interference* (*MUI*), and *inter-cell CCI* or simply *inter-cell interference* (*ICI*), respectively [23]. Fig. 1.7 provides a graphical representation of the various "flavors" of CCI.

Noise and interference limit the data rate and may result in high *bit error rate* (*BER*). Therefore, they can significantly degrade the performance, quality, and reliability of communication (e.g., resulting in high download latency, low speech and video quality, and service interruptions, respectively).

The impact of noise on a radio link is commonly represented by the *receive signal-to-noise-ratio* (*SNR*), i.e., the ratio of the receive power P_r over the noise power N [23]:

$$\text{SNR} = \frac{P_r}{N}. \tag{1.1}$$

Similarly, the *receive signal-to-interference-ratio* (*SIR*)

$$\text{SIR} = \frac{P_r}{I} = \frac{P_r}{\sum_{n=1}^{N} I_n}, \tag{1.2}$$

models the effect of interference. The total interference power I in Eq. (1.2) is the sum of the powers I_n of all interference components at the receiver ($n = 1, \ldots, N$) [23].

The combined impact of noise and interference is expressed via the *receive signal-to-interference-plus-noise-ratio* (*SINR*) [23]:

$$\text{SINR} = \frac{P_r}{I + N}. \tag{1.3}$$

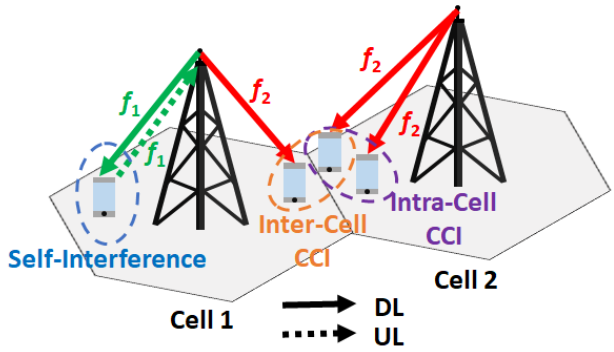


Figure 1.7 The various types of co-channel interference: UL / DL self-interference, intra-cell, and inter-cell.

1.4.2 Large-Scale and Small-Scale Fading

Mobile radio communication encounters several complications: The IOs attenuate the transmitted radio signal and enable propagation over multiple paths, which can lead to distortion. Furthermore, the radio propagation characteristics change with the location as well as with time, due to the movement of the MS or / and the IOs. As a consequence, the receive power fluctuates randomly over time, space, and frequency. This phenomenon is called *fading*.

The mobile radio channel is described by propagation models. There are two basic types of such models [20–23]:

- *Large-scale propagation models* describe the radio propagation characteristics over large areas (with a radius of hundreds of wavelengths). They focus on the description of the receive power or, equivalently, of the *path loss* (i.e., attenuation), as a function of the propagation distance. Link budgeting and cell planning are based on such models.

- *Small-scale propagation models* describe the radio propagation characteristics over small areas (with a radius in the order of a wavelength) or short time intervals (with a duration in the order of a second). They characterize the rapid and severe random fluctuations of the instantaneous receive power around its average value. These variations are attributed to the time-varying constructive and destructive self-interference of multipath components that reach the MS via reflection, scattering, and diffraction. Small-scale propagation models influence the design of physical-layer communication techniques and medium access control protocols, such as modulation schemes, channel equalization strategies, and user scheduling algorithms.

Fig. 1.8 illustrates the receive power vs. the propagation distance on a log-log scale for different radio propagation models. The blue curve in Fig. 1.8 corresponds to a *large-scale path loss* model. The path loss (and, consequently, the receive power) depends on the propagation distance and the radio environment. The former dependency captures the attenuation caused by the geometric spreading of the radiated energy as the transmitted signal propagates through the free space, in conjunction with the inability of the receive antenna to collect the total transferred energy. The latter dependency reflects the energy dissipation by the IOs[2]. The radio environment is in general different at different locations, even when these locations are equidistant from the BS (i.e., it has different geometry / path profile). This implies that the power loss attributed to the IOs and, therefore, the overall path loss is a random variable. Large-scale path loss models describe the *area-mean receive power* (or, equivalently, the *mean path loss*) as a function of the propagation distance. The area-mean receive power is measured in practice by averaging the random values of the receive power for each BS-MS separation distance in an area with a radius of tens or hundreds of meters (i.e., hundreds of wavelengths), such as a cell, to remove the effect of shadow fading and small-scale fading (which are described next). These models take the form of a power-law formula which states that the area-mean receive power drops (or, equivalently, that the mean path loss grows) exponentially with the propagation distance, as shown in Fig. 1.8. The rate at which the area-mean receive power decays is usually determined by the *path loss exponent* (*PLE*), which depends on the environment, the carrier frequency, and the antenna characteristics (height and gain) and is commonly derived from measurements. Typically, the PLE ζ ranges in 1.5–6, where $\zeta = 2$ corresponds to free-space propagation and $\zeta < 2$ occurs under strong waveguiding (e.g., at long avenues with tall buildings at both sides of the road or in tunnels) [21].

The red curve in Fig. 1.8 corresponds to a *large-scale fading* model, which is also called *long-term fading* or *macroscopic fading* model. This model describes the slow random variations of the *local-mean receive power* around the area-mean receive power (or, equivalently, of the path loss around its mean value) that are caused by the variations in the IOs-induced attenuation (mainly due to shadowing and reflection) as the MS moves over a large area. This large-scale propagation phenomenon is called *shadow fading*. The local-mean receive power is measured in practice by averaging the random values of the receive power for each BS-MS separation distance in an area with a radius of a few tens of wavelengths[3], to remove the effect of small-scale fading. These variations of the path loss are modeled as a zero-mean log-normal (i.e., Gaussian on logarithmic scale) random variable that has a certain standard deviation (given in dB) and is added to the distance-

[2]Note that these dependencies are coupled, since in larger distances there is typically a larger number of IOs [23].

[3]An area with a radius of 10–40 wavelengths is commonly considered, which corresponds to a radius of about 1–10 m for a carrier frequency of 1–3 GHz.

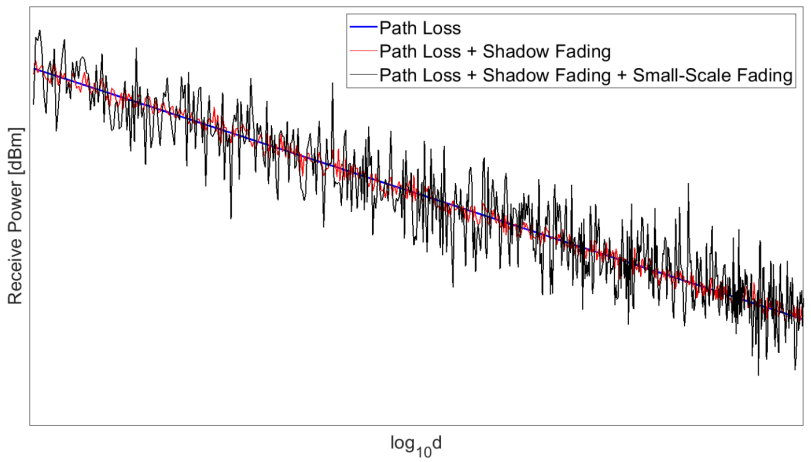


Figure 1.8 Path loss, large-scale fading, and small-scale fading propagation models.

dependent mean path loss component (also given in dB). The standard deviation of the shadow fading component depends on the environment, the carrier frequency, and the antenna characteristics (height and gain) and is commonly derived from measurements. Typically, it ranges in 6–12 dB.

Finally, the black curve in Fig. 1.8 corresponds to a *small-scale fading* model, which is also known as a *short-term fading* or *microscopic fading* model. This model describes the rapid and severe random fluctuations of the *instantaneous receive power* around its local mean (up to 30–40 dB) as the MS or / and the IOs move(s) over a short distance (in the order of a wavelength). Let us clarify the reason why this happens. The transmitted radio signal reaches the MS over multiple NLOS paths (and possibly a LOS path) due to its reflection, diffraction, and scattering by obstacles in the environment. This phenomenon is called *multipath propagation*. These paths have different lengths and induce different attenuations, delays, and phase shifts to the *multipath components* (*MPC*). The MPCs are added vectorially at the MS either constructively or destructively, depending on their phases which, in turn, depend on the path lengths—that is, on the positions of the MS and the IOs [12, 20–23]. Small movements of the MS or / and the IOs (and, therefore, small path length changes) can alter dramatically the phase shifts of the MPCs. More specifically, a spatial displacement of half wavelength can cause a change from constructive to destructive self-interference and vice versa[4] [12, 20]. As a consequence, we notice substantial random variations of the instantaneous receive power over short time intervals (in the order of a second) or small areas (in the order of a wavelength), which are attributed to the movement of the MS or / and the scatterers [12, 20–23]. Note that these movements affect significantly the

[4] A carrier frequency of 2 GHz corresponds to a wavelength of about 15 cm.

receive power, even though the path loss and shadow fading remain essentially constant over such short distances [6]. This phenomenon is called small-scale fading and is typically characterized via the impulse response of the channel, which is commonly represented by a complex Gaussian process[5].

In summary:

- Large-scale fading models describe the slow random variations of the local-mean receive power around the area-mean receive power as the MS moves over a large area (hundreds of meters in size). They incorporate the effects of distance-dependent path loss and shadow fading.

- Small-scale fading models describe the rapid and severe random variations of the instantaneous receive power around the local-mean receive power over short time scales (in the order of a second) or for small spatial displacements of the MS or / and the IOs (in the order of a wavelength). This phenomenon is attributed to multipath propagation and to the motion of the MS or / and the obstacles in the surrounding environment.

We should note that often large-scale fading is referred to as shadowing and small-scale fading is simply called fading in the literature.

1.5 Large-Scale and Small-Scale Fading Models

In this section, we present the main models considered for the description of the large-scale fading and small-scale fading phenomena.

1.5.1 Large-Scale Fading Models

Large-scale propagation is described by the mean path loss and the slow variations of the path loss around its mean value due to shadowing.

Some simple models for describing the distance-dependent path loss include the *free space path loss* (*FSPL*) model, which considers only a LOS path, and the *ground-reflection* (*two-rays*) *model*, which takes also into account a ground-reflection component [20, 23]. The *log-distance path loss model* is a more evolved model, in that it describes the effect in radio propagation of the various IOs in the environment as well. It is a quite popular model because it is simple and mathematically tractable, yet both generic enough and reasonably accurate for most intends and purposes. The log-distance path loss model corresponds to the blue curve in Fig. 1.8.

[5]This reflects a highly cluttered environment and it is a consequence of the *central limit theorem* (*CLT*) [23, 25]. In the absence of a LOS path, the mean of this Gaussian process is zero and its envelope is *Rayleigh* distributed.

According to this model, the area-mean receive power at distance d meters from the BS is given in linear units (e.g., in watts) by [23]:

$$\overline{\overline{P}}_r(d) = P_t \overline{L}_0 \left(\frac{d}{d_0} \right)^{-\zeta}, \tag{1.4}$$

where P_t is the transmit power in linear units. That is [23],

$$\overline{\overline{P}}_r(d)|_{\text{dBm}} = P_t|_{\text{dBm}} + \overline{L}_0|_{\text{dB}} - 10\zeta \log_{10} \frac{d}{d_0}, \tag{1.5}$$

where $\overline{L}_0|_{\text{dB}} = 10 \log_{10} \overline{L}_0$.

The linear mean path loss at distance d from the BS is given by the ratio of the transmit power over the area-mean receive power at this distance [23]:

$$\overline{L}_p(d) = -\overline{L}_0 \left(\frac{d}{d_0} \right)^{-\zeta}. \tag{1.6}$$

Similarly, the mean path loss is expressed in dB as the difference between the transmit power and the area-mean receive power, when these quantities are expressed in logarithmic units (e.g., in dBm). That is [23],

$$\overline{L}_p(d)|_{\text{dB}} = -\overline{L}_0|_{\text{dB}} + 10\zeta \log_{10} \frac{d}{d_0}. \tag{1.7}$$

Clearly, the log-distance path loss model is fully characterized by three parameters:

- the path loss exponent (PLE) ζ;
- the reference distance d_0; and
- the mean path loss at the reference distance $\overline{L}_0$.

The PLE determines how fast the area-mean power decays with the propagation distance. We note that $\overline{\overline{P}}_r(d)$ drops with the ζ-th power of d (i.e., with a rate of 10ζ dB/decade). The PLE depends on the environment, the carrier frequency, and the antenna characteristics (height and gain). It is typically derived from measurements and ranges between 1.5 and 6 [20] (e.g., for NLOS macrocellular setups operating at 2 GHz, $\zeta = 3.76$ [26]). For $\zeta = 2$ the log-distance path loss model approximates the FSPL model, whereas for $\zeta = 4$ it approximates the ground reflection (two-rays) path loss model for large distances, where the LOS component and the ground-reflected component are added destructively. Values of $\zeta < 2$ can be observed in urban areas due to waveguiding in "street canyons" [21].

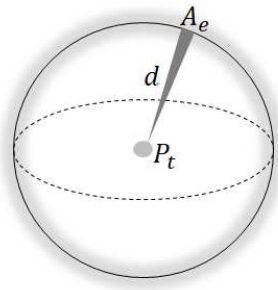


Figure 1.9 The radiated energy is spread uniformly on the surface of an ever-expanding sphere, but the receive antenna can capture only a small fraction of it, which is determined by its effective area.

The *reference distance* $d_0 < d$ is in the far-field of the BS[6]. For indoor small cells, we typically set $d_0 = 1$ m, while for outdoor small cells we set $d_0 = 10$ m or $d_0 = 100$ m [20, 23]. Similarly, for macro cells we set $d_0 = 1$ km [20].

The mean path loss at the reference distance d_0, $\overline{L}_0$, depends on the environment, the carrier frequency, and the antenna characteristics (height and gain). It is often derived from measurements (typically in dB) by averaging the random values of the receive power at distance d_0 from the BS over an area of tens or hundreds of meters, to remove the effect of shadow fading and small-scale fading. For instance, for NLOS macrocellular setups operating at 2 GHz, the mean path loss at a reference distance of 1 km is 148.1 dB [26]. As an alternative, $\overline{L}_0$ can be expressed as the FSPL at distance d_0 assuming the use of omnidirectional antennas:

$$\overline{L}_0|_{\text{dB}} = -20\log_{10}\frac{\lambda}{4\pi d_0} = -20\log_{10}\frac{c}{4\pi d_0 f_c}. \tag{1.8}$$

In Eq. (1.8), $c = 3 \times 10^8$ m/s is the speed of light in vacuum, λ is the wavelength in meters, f_c is the carrier frequency in Hertz, and $\lambda = c/f_c$. We note that in a free space propagation environment the transmitted radio signal is attenuated with the square of the propagation distance, as mentioned earlier, and the square of the carrier frequency. The distance-dependence reflects the fact that the radiated energy is spread uniformly on the surface of an ever-expanding sphere, while the frequency-dependence reflects the fact that the receiver captures only a fraction of the radiated energy that is determined by its antenna's *aperture* or *effective area*, A_e, which is a frequency-dependent quantity (see Fig. 1.9).

Shadow fading refers to the slow random variations of the receive power around its area-mean value, which are caused by the changes in the radio environment

[6]The far-field region starts at the *Fraunhofer distance* $d_f = 2D^2/\lambda$, where D is the largest linear dimension of the transmit and receive antennas and λ is the wavelength [20].

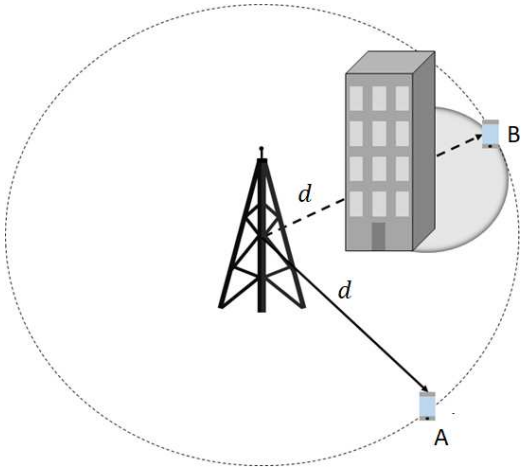


Figure 1.10 Locations A and B are equidistant from the BS. However, at location A communication takes place over a LOS path, whereas at location B the MS is in the shadow area of a large building.

and, therefore, in shadowing (and energy absorption by large obstacles in general) as the MS moves over a large area, as depicted in Fig. 1.10. This local-mean of the receive power, $\overline{P}_r(d)$, is the enseble average of the receive power for a fixed BS-MS separation distance d calculated over an area with a radius of a few tens of wavelengths (typically 10–40), to remove the effect of small-scale fading [20].

These variations are modeled as a zero-mean Gaussian random variable expressed in dB, $X_{sf}|_{\text{dB}} \sim \mathcal{N}\left(0, \sigma_{sf}^2\right)$, that is added to the mean path loss (also expressed in dB) [20, 23]:

$$L_p(d)|_{\text{dB}} = \overline{L}_p(d)|_{\text{dB}} + X_{sf}|_{\text{dB}}. \tag{1.9}$$

The local-mean receive power is given by:

$$\begin{aligned} \overline{P}_r(d)|_{\text{dBm}} &= P_t|_{\text{dBm}} - L_p(d)|_{\text{dB}} \\ &= P_t|_{\text{dBm}} - \overline{L}_p(d)|_{\text{dB}} - X_{sf}|_{\text{dB}} \\ &= \overline{\overline{P}}_r(d)|_{\text{dBm}} - X_{sf}|_{\text{dB}}. \end{aligned} \tag{1.10}$$

This model, which combines the log-distance path loss with the log-normal variations of the path loss around its mean value caused by shadowing, is called the *log-normal shadowing model* and describes the large-scale fading variations. The log-normal shadowing model corresponds to the red curve in Fig.1.8.

Note that the reason why the mean value of $X_{sf}|_{\text{dB}}$ is assumed to be zero is because the mean path loss $\overline{L}_p(d)|_{\text{dB}}$ is included explicitly in these equations [23].

Otherwise, the mean value of $X_{sf}|_{\text{dB}}$ would equal the mean path loss (in dB). Notice also that $X_{sf}|_{\text{dB}}$ is specified in terms of its standard deviation, σ_{sf}, which depends on the environment, the carrier frequency, and the antenna characteristics (height and gain) and is commonly derived from measurements. Typically, σ_{sf} takes values in 6–12 dB [6] (e.g., for NLOS macrocellular setups operating at 2 GHz, $\sigma_{sf} = 10$ dB [26]).

Finally, we should mention that the linear path loss $L_p(d)$ is a non-negative quantity. The linear *path gain*, $G_p(d)$, is the inverse of the linear path loss—thus, the path gain in dB is the negative of the path loss in dB.

1.5.2 Small-Scale Fading Models

Small-scale fading is the combined effect of two phenomena, namely, the multipath propagation and the movement of the MS or / and the IOs in the surrounding environment. It refers to the substantial variations in the constructive and destructive self-interference of the received MPCs caused by the dramatic alterations of their phase shifts due to the small changes in the path lengths resulting from small spatial displacements of the MS or / and the scatterers.

A fading channel is modeled as a *linear time-variant* (*LTV*) *system*[7]. Therefore, it is described by the *channel impulse response* (*CIR*), which is, in general, a function of the absolute time t and the propagation delay τ, $h(t,\tau)$. Equivalently, it can be described by the corresponding *channel transfer function* (*CTF*), $H(t,f)$[8].

Multipath propagation determines the filtering effect of the channel, that is, its frequency-selectivity / time-dispersion properties. Let us elaborate this argument, starting from a description of this concept in the frequency domain (see Fig. 1.11). The *coherence bandwidth* of the channel, B_c, is the frequency range over which the *channel frequency response* (*CFR*) $H(f)$ (i.e., the CTF $H(t,f)$ at a given time instant) is roughly constant. Therefore, two sinusoids with a frequency separation smaller than the coherence bandwidth are correlated. If the signal bandwidth is larger than the coherence bandwidth (i.e., if $B > B_c$), the channel affects differently the different spectral components of the input signal, thus leading to *multipath distortion*. We say that this is a *frequency-selective fading channel* or a *wideband fading channel*. Otherwise (i.e., if $B \ll B_c$), the channel passes all the frequency components of the

[7] A fading channel is modeled as a system, since it responds to an applied input (the transmitted signal) by producing an output (the received signal). This system is linear, because it corresponds to a linear operation, that is, to the superposition of the *multipath components* (*MPC*) at the receiver. Also, it is time-variant due to the motion of the MS or / and the IOs.

[8] Under the *wide sense stationary uncorrelated scattering* (*WSSUS*) model, we can describe the channel via certain *correlation functions* and *power spectral density functions*, such as the *scattering function*, the *power delay spectrum*, the *Doppler spectrum*, the *time correlation function*, and the *frequency correlation function*. These *stochastic system functions*, in turn, enable us to formally define parameters that characterize the behavior of the channel, such as the coherence bandwidth and the coherence time. The description of these functions as well as of the relations between them are beyond the scope of this introductory chapter. The interested reader may refer to standard textbooks, such as [12, 20, 21, 23], for more details.

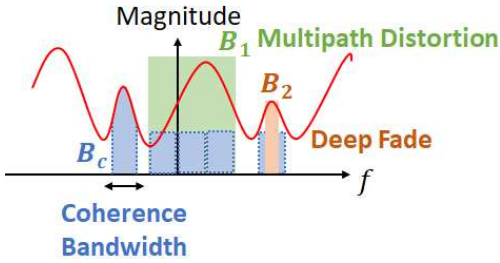


Figure 1.11 Multipath propagation is responsible for the frequency-selectivity of the channel.

input signal with the same gain and introduces a linear phase shift, thus resulting in distortionless transmission. Such a channel is called a *frequency-flat fading channel* or a *narrowband fading channel*. Flat-fading channels result in SNR loss when a deep fade occurs, as the channel varies over time due to the motion of the MS or / and the scatterers. Let us now describe the same phenomenon in the delay domain (see Fig. 1.12). The maximum delay spread T_m of the channel is the time difference between the delay of the first arriving MPC at the MS (typically the LOS component, if there is one) and the last one[9]. If the symbol period T is smaller than the maximum delay spread (i.e., if $T < T_m$), which is often the case for high data rate communication systems, then multiple MPCs are resolvable and the channel induces significant time-dispersion to the transmitted signal due to the different delays of the MPCs[10]. This results in *inter-symbol interference* (*ISI*) and, therefore, in high BER. Otherwise (i.e., if $T \gg T_m$), the individual MPCs are not resolvable and the introduced time-dispersion is negligible. Such a channel simply attenuates the transmitted signal[11]. The coherence bandwidth and the maximum delay spread (or the RMS delay spread) are inversely proportional to each other. Note that in the same radio environment two different systems with different symbol period / signal bandwidth might encounter different types of small-scale fading, regarding the frequency-selectivity of the channel.

The mobility of the MS or / and the IOs determines the time-variability effect of the channel, i.e., its time-selectivity / frequency-dispersion properties. Let us describe this concept in the time domain. The *coherence time* of the channel, T_c, is the time interval over which the CIR remains roughly constant. If the symbol

[9] In practice, we consider the last MPC whose power is above a predefined threshold that is determined by the background noise level and the sensitivity of the receiver [23].

[10] A resolvable path may be associated with a single scatterer or it may correspond to multiple paths with similar delays associated with a cluster of scatterers. In the latter case, the attenuation and phase shift of the resolvable MPC results from the combination of the individual MPCs that consist this path [23].

[11] In practice, we typically compare the symbol period T with the root-mean-square (RMS) value of T_m, that is, with the *RMS delay spread* τ_m [23].

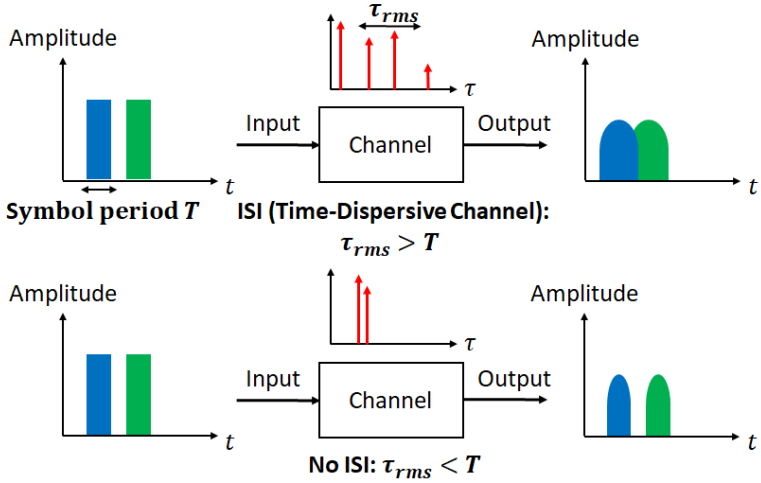


Figure 1.12 Time-dispersion of the transmitted signal due to multipath may result in inter-symbol interference (ISI).

period is larger than the coherence time (i.e., if $T > T_c$), then the channel characteristics change multiple times during a transmission slot. Such a channel is referred to as a *fast fading channel*. Otherwise (i.e., if $T \ll T_c$), the channel remains roughly constant during a transmission slot. Such a channel is called a *slow fading channel*. Fast fading poses difficulties for synchronization and channel estimation which, in turn, may result in performance degradation (e.g., precoding may be based on inaccurate channel knowledge or we may have to rely on statistical channel knowledge, which leads to performance loss). Let us now describe this concept in the Doppler domain. The different Doppler (i.e., frequency) shifts of the MPCs due to motion result in spectral broadening of the transmitted signal. If the signal bandwidth is smaller than the resulting *Doppler spread* (i.e., if $B < B_D$), then the encountered spectral broadening is large; otherwise (i.e., if $B \gg B_D$), it is negligible. The coherence time and Doppler spread are inversely proportional to each other. A special case of interest is a *block fading channel*, where the impulse response remains roughly constant for the duration of a transmission block of N symbols. In this scenario, $T_c \gg N_T = NT$. We should mention that the same time-variant multipath channel may appear as fast varying for one radio communication system and as slowly varying for another, depending on the relation of the Doppler spread or coherence time with the signal bandwidth or symbol period, respectively. Clearly, very low data rate communication systems suffer more from the time-selectivity of the mobile radio channel. This is only half of the story, though: The Doppler spread / the temporal variance of the channel is a major challenge when the users are moving fast (e.g., passengers in a train) or the carrier frequency is high.

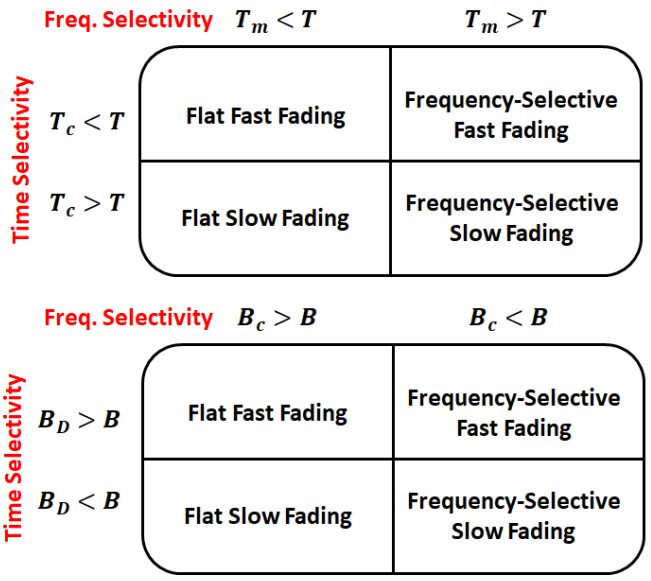


Figure 1.13 Classification of fading channels.

Based on the above discussion, it becomes apparent that we can classify the fading channels into four categories according to their frequency-selectivity and time-selectivity, as depicted in Fig. 1.13.

Let's assume a fast fading channel. If this channel exhibits frequency-selective fading (i.e., it is wideband), then the output is given by the convolution of the input with the impulse response (ignoring the additive noise):

$$y(t) = x(t) * h(t,\tau). \tag{1.11}$$

The channel impulse response is given by:

$$h(t,\tau) = \sum_{n=1}^{N} \tilde{\alpha}_n(t)\delta\left(\tau - \tau_n(t)\right) = \sum_{n=1}^{N} \alpha_n(t)e^{j\theta_n(t)}\delta\left(\tau - \tau_n(t)\right), \tag{1.12}$$

where $\tilde{\alpha}_n(t) = \alpha_n(t)e^{j\theta_n(t)}$. We note that the channel impulse response is expressed as the superposition of N MPCs with different amplitudes $\alpha_n(t)$, phases $\theta_n(t)$, and delays $\tau_n(t)$ due to the wideband channel assumption, and these parameters are time-varying due to the fast fading assumption.

If, on the other hand, the channel exhibits frequency-flat fading (i.e., it is narrowband), then the output equals the product of the input with the impulse response:

$$y(t) = x(t)h(t). \tag{1.13}$$

Note that in this case the impulse response is simply described as [24]

$$h(t) = \tilde{\alpha}(t) = \alpha(t)e^{j\theta(t)}. \tag{1.14}$$

That is, the channel impulse response reduces to a single impulse.

Now, if we assume that the channel is quasi-static (i.e., time-invariant over a transmission slot or block), then for the case of a wideband channel we have:

$$h(\tau) = \sum_{n=1}^{N} \tilde{\alpha}_n \delta(\tau - \tau_n) = \sum_{n=1}^{N} \alpha_n e^{j\theta_n} \delta(\tau - \tau_n). \tag{1.15}$$

Note that we have dropped the time index, since the attenuations, phases, and delays are time-invariant over the time window of interest.

Similarly, for a narrowband quasi-static channel, we have:

$$h = \tilde{\alpha} = \alpha e^{j\theta}. \tag{1.16}$$

That is, during the transmission slot or block, the channel is modeled as a complex scalar.

Note that the large-scale and small-scale radio propagation phenomena are mutually independent and multiplicative. That is, if, let's say, the small-scale fading coefficient is a zero-mean complex Gaussian process with unit variance $h \sim \mathcal{CN}(0,1)$, then the wireless channel can be modeled as $\sqrt{G_p}h$, where G_p is the path gain[12]. Equivalently, we could simply model the channel as $h \sim \mathcal{CN}(0, G_p)$. We should note, though, that we often assume that the path loss and shadowing have been compensated via power control / adaptive power allocation schemes and focus only on the small-scale fading effects.

1.6 Performance under Fading, Noise, and Interference

The available communication resources, the noise and interference levels, and the fading degradations determine the performance of cellular networks.

1.6.1 Performance Metrics

The performance of a radio link is assessed via the *capacity* of the wireless channel, i.e., the maximum data rate (in bits/s) that can be supported by the channel with arbitrarily low probability of error (under some idealizations) [8]. Since the spectrum is a scarce resource, we are also interested in the *spectral efficiency* (*SE*). This performance metric refers to the capacity per unit of bandwidth and it is, naturally,

[12]Note that the real and imaginary part of h are zero-mean Gaussian processes with variance $1/2$.

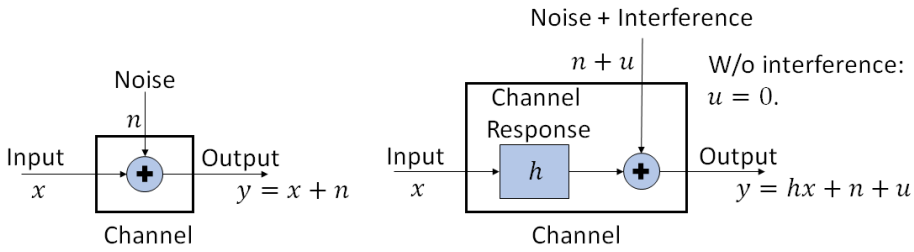


Figure 1.14 The additive noise (left), flat-fading with additive noise (right with $u = 0$), and flat-fading with additive noise and interference (right with $u \neq 0$) channels.

measured in bits/s/Hz [8]. Clearly, higher SE implies more efficient utilization of the available spectrum.

At cell or network level, we quantify the performance via the *sum-rate* (*SR*) *capacity* and the *sum-SE*, which are defined as the sum of the capacities or SEs, respectively, of the respective links.

1.6.2 AWGN Channel

Let us consider a radio link between a BS and a MS and assume initially that it corresponds to a fictional wireless channel that simply contaminates the received signal with AWGN. As illustrated in Fig. 1.14, the complex-baseband representation of the received signal in this case is:

$$y = x + n, \tag{1.17}$$

where $x \in \mathbb{C}$ and $y \in \mathbb{C}$ are the transmitted and received signal, respectively, and $n \sim \mathcal{CN}(0, N)$ is additive zero-mean Gaussian noise with *power spectral density* (*PSD*) $N_0/2$ W/Hz and variance $N = N_0 B$ watts. The channel input x is a zero-mean stochastic process as well. Communication is subject to an average transmission power constraint $\mathbb{E}\left\{|x|^2\right\} \leq P_t$.

The capacity of this AWGN channel is given by [25]

$$\begin{aligned} C(P_t, B) &= B \log_2\left(1 + \frac{P_r}{N}\right) \\ &= B \log_2\left(1 + \frac{P_t}{N_0 B}\right) \\ &= B \log_2(1 + \mathrm{SNR}), \end{aligned} \tag{1.18}$$

where $P_r = P_t$ is the receive power[13] and

$$\text{SNR} = \frac{P_r}{N} = \frac{P_t}{N_0 B} \tag{1.19}$$

is the receive SNR. The capacity-achieving input is $x \sim \mathcal{CN}(0, P_t)$.

In the low SNR regime, where $\text{SNR} \ll 1$, we obtain [25]:

$$\begin{aligned} C(P_t, B) &= B \log_2 \left(1 + \frac{P_r}{N}\right) \\ &\approx B \log_2 \left(\frac{P_r}{N}\right) \log_2 e \\ &= B \left(\frac{P_t}{N_0 B}\right) \log_2 e \\ &= \frac{P_t}{N_0} \log_2 e. \end{aligned} \tag{1.20}$$

We notice that the capacity grows linearly with P_t and is insensitive in B. This is called the *power-limited* SNR regime.

In the high SNR regime, where $\text{SNR} \gg 1$, we have [25]:

$$\begin{aligned} C(P_t, B) &= B \log_2 \left(1 + \frac{P_r}{N}\right) \\ &\approx B \log_2 \left(\frac{P_r}{N}\right) \\ &= B \log_2 \left(\frac{P_t}{N_0 B}\right). \end{aligned} \tag{1.21}$$

We see that the capacity grows logarithmically with P_t and approximately linearly[14] with B. This is known as the *bandwidth-limited* SNR regime. We note that in this SNR regime, the system bandwidth is a more valuable resource than the transmit power, from a capacity-enhancement perspective. Indeed, doubling B almost doubles the capacity, while doubling P_t increases the capacity by only 1 bit/s.

[13] There is no reason to transmit with smaller than the maximum allowable power.

[14] The large linear increase of the capacity with B in the high SNR regime compensates for the small logarithmic performance loss due to the increase of $N = N_0 B$. However, as B (and, therefore, N) increases without bound and P_t is fixed, the system drops to the low SNR regime and the capacity converges to the approximation given by Eq. (1.20).

1.6.3 Frequency-Flat Block Fading Channel

Now, let us consider a frequency-flat block fading channel. The complex-baseband representation of the channel output is [27]

$$y = hx + n. \tag{1.22}$$

The channel is typically modeled as a Gaussian process. We assume that the channel realization h is known at the MS. In other words, we assume that *channel state information* (*CSI*) is available at the receiving side (*CSIR*). Such knowledge can be obtained through pilot-assisted channel estimation.

The *instantaneous capacity* for a given channel realization is [27]

$$\begin{aligned} C(P_t, B) &= B \log_2 \left(1 + \frac{P_r}{N}\right) \\ &= B \log_2 \left(1 + \frac{|h|^2 P_t}{N_0 B}\right) \\ &= B \log_2 \left(1 + |h|^2 \mathrm{SNR}\right) \\ &= B \log_2 (1 + \gamma), \end{aligned} \tag{1.23}$$

where $|h|^2$ denotes the *instantaneous channel gain* and

$$\gamma = \frac{P_r}{N} = \frac{|h|^2 P_t}{N_0 B} = |h|^2 \mathrm{SNR} \tag{1.24}$$

is the *instantaneous receive SNR*, i.e., the ratio of the instantaneous receive power $P_r = |h|^2 P_t$ over the noise power $N = N_0 B$.

Since the instantaneous receive power and, therefore, the instantaneous receive SNR fluctuate over time due to the randomness of the channel response, it is more meaningful to determine the *ergodic capacity* of the channel [27]:

$$\begin{aligned} \overline{C}(P_t, B) &= \mathbb{E}\{C\} \\ &= \mathbb{E}\left\{B \log_2 \left(1 + \frac{P_r}{N}\right)\right\} \\ &= \mathbb{E}\left\{B \log_2 \left(1 + \frac{|h|^2 P}{N_0 B}\right)\right\} \\ &= \mathbb{E}\left\{B \log_2 \left(1 + |h|^2 \mathrm{SNR}\right)\right\} \\ &= \mathbb{E}\{B \log_2 (1 + \gamma)\} \\ &= B \log_2 (1 + \overline{\gamma}), \end{aligned} \tag{1.25}$$

where $\mathbb{E}\left\{|h|^2\right\}$ denotes the *average channel gain* and

$$\begin{aligned}\overline{\gamma} &= \mathbb{E}\{\gamma\} \\ &= \mathbb{E}\left\{\frac{P_r}{N}\right\} \\ &= \mathbb{E}\left\{|h|^2 \mathrm{SNR}\right\} \\ &= \mathbb{E}\left\{|h|^2\right\} \mathrm{SNR} \\ &= \mathbb{E}\left\{|h|^2\right\} \left(\frac{P_t}{N_0 B}\right)\end{aligned} \tag{1.26}$$

is the *average receive SNR*, i.e., the ratio of the average receive power $\overline{P}_r = \mathbb{E}\{P_r\} = \mathbb{E}\left\{|h|^2\right\} P_t$ over the noise power $N = N_0 B$. Notice that the expectation is taken w.r.t. h.

1.6.4 Frequency-Flat Block Fading Channel with Interference

So far, we have implicitly assumed an isolated radio link. If we consider also the interference by other transmissions in the cellular network, we obtain [27] (see Fig. 1.14)

$$y = hx + u + n, \tag{1.27}$$

where $u \in \mathbb{C}$ denotes the total interference that affects the intended transmission. The interference is commonly modeled as a zero-mean additive stochastic process u that is independent from the noise and uncorrelated with the input. Assuming that the interference variance I is known at the MS, the instantaneous capacity of the channel is lower bounded as [27]

$$\begin{aligned}C(P_t, B) &\geq B \log_2\left(1 + \frac{P_r}{I + N}\right) \\ &= B \log_2\left(1 + \frac{|h|^2 P_t}{I + N_0 B}\right) \\ &= B \log_2(1 + \mathrm{SINR}),\end{aligned} \tag{1.28}$$

where

$$\mathrm{SINR} = \frac{P_r}{I + N} = \frac{|h|^2 P_t}{I + N_0 B} \tag{1.29}$$

is the *instantaneous receive SINR*, i.e., the ratio of the instantaneous receive power $P_r = |h|^2 P_t$ over the sum of the interference power I and the noise power $N = N_0 B$. Similarly, the ergodic capacity of this channel is lower bounded as [27]

$$\overline{C}(P_t, B) \geq \mathbb{E}\{B \log_2(1 + \mathrm{SINR})\}. \tag{1.30}$$

The expectation is taken w.r.t. both h and u. In both cases, the bound is achieved by the input $x \sim \mathcal{CN}(0, P_t)$.

Notice that these bounds assume that we treat the randomly varying interference as additional noise. While in the low interference regime this strategy is indeed optimal, in the high interference regime it would be better to explicitly mitigate or eliminate the interference [27].

1.7 Noise, Shadowing, and Fading Mitigation

The contamination of the received signal by thermal background noise is unavoidable in a communication system. Also, fading is a characteristic of mobile radio communications and arises as the MS moves over both large and small areas (although with a different form and as a result of different mechanisms). In this section, we present noise, shadowing, and (small-scale) fading mitigation techniques.

1.7.1 Noise Mitigation Techniques

Several noise mitigation techniques have been studied in the literature [24]:

- Some *modulation formats* are more robust against noise in comparison to others (i.e., modulated signals that do not imprint the information signal onto their amplitude variations and lower-order modulation schemes, which have a less "crowded" constellation diagram).
- *Forward error correction* (*FEC*) (i.e., *channel coding*) schemes introduce redundancy in a controlled manner to enable the detection and correction of errors, thus reducing the required SNR for reliable communication (i.e., they provide a *coding gain*).
- *Matched filtering* maximizes the SNR at the output of the receive filter.
- When multiple antennas are utilized at the TX or / and the RX, *transmit / receive beamforming* (*BF*) can be utilized to focus spatially the corresponding antenna radiation pattern(s) and, therefore, increase the effective transmit / receive power and the receive SNR (i.e., BF provides *array gain / BF gain*).

1.7.2 Shadowing Mitigation Techniques

Shadow fading may drop the receive SNR below the minimum required value for acceptable performance, thus resulting in *outage*. Common approaches to overcome this issue include the dynamic adjustment of the transmit power and the addition of a *fading margin* to the transmission power [20]. However, care should be taken in both cases to avoid the occurrence of harmful interference. Another

strategy is *macroscopic diversity*, where the signals from two BSs are combined at the MS [20]. This method, though, requires cooperation between the BSs.

1.7.3 Fading Mitigation Techniques

Small-scale fading is manifested in several ways. Different mitigation techniques are used for each type of fading degradations [22].

- Common mitigation strategies for frequency-selective fading include *adaptive channel equalization*, to remove the ISI from the received signal, and *orthogonal frequency division multiplexing* (*OFDM*), to convert the frequency-selective channel into a number of parallel frequency-flat sub-channels. A common approach for mitigating the effect of flat fading, on the other hand, is to use channel coding in combination with *diversity* techniques that provide uncorrelated estimations of the transmitted signal at the receiver (e.g., *frequency diversity* via *sub-carrier interleaving* in OFDMA systems, *transmit / receive spatial diversity* in multi-antenna communication systems, etc.).

- Error correction coding and *bit-interleaving*, which introduces time diversity to avoid burst errors; robust *non-phase-coherent* or *differential-phase modulation schemes*, which do not require phase synchronization; and *redundancy techniques* that increase the transmission rate are commonly utilized to mitigate the effects of fast fading. For slow fading, *block-interleaving* may be used, in conjunction with channel coding.

1.8 CCI Management Techniques

The shared nature of the wireless channel leads to the occurrence of interference. In this section, we present co-channel interference (CCI) management methods.

1.8.1 Duplexing Methods

From a SE perspective, a non-orthogonal duplexing strategy, wherein a BS (MS) transmits signals in the DL (UL) and receive signals in the UL (DL) simultaneously over the same frequency band, would be optimal. However, in such an *in-band full duplex* (*IBFD*) or simply *FD* system one would have to deal with a staggering amount of more than 100 dB of self-interference (SI). The high level of SI is attributed to the proximity of the transmit and receive antenna ports. Several SI suppression techniques have been studied in the last few years, and some of them seem really promising (e.g., see [28] and references therein). Nevertheless, there are many implementation issues that have to be addressed prior to the commercial application of this technology.

Therefore, cellular networks rely on *duplexing* schemes that orthogonalize the DL and UL transmissions in the frequency or time domain, in order to enable bi-directional communication. More specifically, in *frequency-division duplex* (*FDD*) or *out-of-band full duplex* (*OBFD*) these transmissions take place on different frequencies, whereas in *time-division duplex* (*TDD*) or *half-duplex* (*HD*) they occur at different times.

The applied duplexing scheme determines the method with which CSI is obtained at the BS. Availability of CSI at the transmitter (*CSIT*) is required, as we shall see, for precoding. In FDD systems, CSIT is obtained via pilot-assisted estimation of the DL channels at the MSs and report of the channel estimates to the BS via a feedback UL channel. TDD systems, on the other hand, rely on pilot-assisted estimation of the UL channels at the BS and exploitation of the reciprocity between the DL and UL channels, since in this case the DL and UL transmissions take place on the same frequency. Thus, in TDD the CSIT acquisition overhead is smaller. TDD suits also better systems that utilize a large number of BS antennas. This is because an UL pilot can be "heard" by all BS antennas and, therefore, the overhead (number of pilots) in TDD operation for learning the DL channels scales with the number of MSs instead of the number of BS antennas. Other advantages of TDD over FDD include the use of unpaired spectrum, which is translated into cost savings for the operators; the utilization of simpler and less expensive transceivers, since TDD does not require transmit/receive filters (diplexers) to avoid the occurrence adjacent channel interference; and its ability to dynamically adjust the number of resources (*timeslots*) allocated to the DL and UL transmissions (i.e., the *DL/UL ratio*), in order to support efficiently *asymmetric DL/UL traffic* [29–31]. On the other hand, TDD has also some drawbacks [29–31]: First of all, it requires calibration of the TX and RX hardware to compensate for mismatches and make the end-to-end effective channel (that includes the TX and RX filters) reciprocal. Also, in TDD operation guard periods should be included between the DL and UL transmissions, since otherwise the UL transmission of a cell-center user may cause interference to the DL reception of a cell-edge user. These guard periods limit the capacity as well as the range, since larger cells require larger guard periods. Furthermore, TDD requires inter-cell synchronization to avoid the ICI caused by an UL transmission in one cell and a DL transmission in an adjacent cell (both associated with cell-edge users). This is not an issue in FDD systems, since the UL and DL transmissions use different frequencies. Finally, since in TDD the transmitter is silent for half of the time in the DL, the receive SNR is 3 dB lower than the one achieved by an equivalent FDD system.

1.8.2 Multiple Access Methods

Multiple access methods divide the available communication resources among the DL transmissions, to enable multiple users to access the shared physical medium.

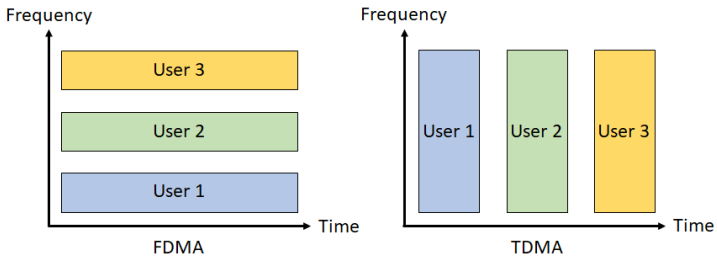


Figure 1.15 FDMA and TDMA techniques.

Frequency-division multiple access (*FDMA*) and *time-division multiple access* (*TDMA*) are orthogonal multiple access schemes wherein the DL transmissions take place on different frequencies or at different times, respectively [20, 23, 24], as highlighted in Fig. 1.15. FDMA and TDMA can be combined into a FDMA/TDMA scheme.

Orthogonal FDMA (*OFDMA*) is a more advanced scheme, wherein the system bandwidth is divided into subchannels [23, 24]. A different group of subchannels is assigned to each user, as shown in Fig. 1.16. The BS communicates with multiple users concurrently over their respective subchannels. OFDMA enables dynamic allocation of the time-frequency resources to the users and provides robustness to channel fading through the conversion of a high rate transmission to parallel low rate transmissions as well as via frequency diversity.

Space-division multiple access (*SDMA*) is a non-orthogonal scheme where a BS equipped with multiple antennas exploits the spatial dimension to serve a group of users on a single time-frequency resource and mitigate or cancel the resulting intra-cell MUI [8]. The signal processing operations that take place prior to transmission to enable the management of the inter-user interference are referred to as *precoding* and require knowledge of the corresponding DL channels. In LOS conditions, precoding resembles multi-stream transmit beamforming (BF): it refers to

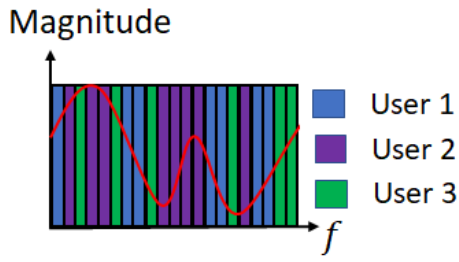


Figure 1.16 OFDMA.

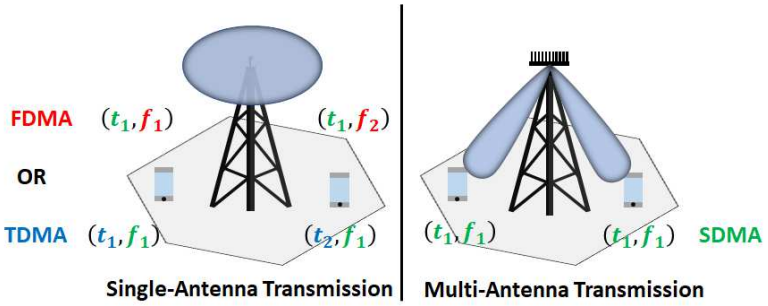


Figure 1.17 Difference between FDMA, TDMA, and SDMA.

the steering of angular beams towards the intended users, so that the interference at non-intended users is decreased [5, 27]. In NLOS conditions, it ensures that the multipath components of each transmitted signal add constructively at the respective user and destructively at other users [5, 27]. This technology is called more formally *multi-user multiple-input multiple-output* (*MU-MIMO*) and is an integral component of contemporary cellular networks. MU-MIMO increases the SE of the cell at the expense of additional signal processing complexity as well as higher cost and power consumption, since each antenna is fed by an RF unit. Fig. 1.17 illustrates the difference between FDMA, TDMA, and SDMA.

1.8.3 ICI Management

Operators exploit the attenuation of the radio signals with the propagation distance to increase the network capacity. More specifically, they re-use the available spectrum at cells that are sufficiently spaced apart, so that the resulting ICI is negligible [20]. This *frequency re-use* concept is applied as follows: Let's say that the system bandwidth is comprised by K channels. These channels are divided into N groups of $S = K/N$ channels. Then, each group is assigned to a cell, so that the whole spectrum is utilized in a cluster of N cells with disjoints channel groups. This cluster is replicated M times over the service area, thus resulting in $C = M \times N \times S = M \times K$ channels. The applied *frequency re-use pattern* ensures that the resulting ICI levels are tolerable. Otherwise, this type of CCI, which is more prominent at the cell boundaries, would degrade the QoS of the cell-edge users and limit the overall SE. Fig. 1.18 shows an example of a frequency re-use pattern with $N = 7$ (i.e., with a *frequency re-use factor* $F_R = 1/N = 1/7$). In practice, sectorization is typically also applied to further reduce the ICI and improve the SNR, thanks to the directional transmissions.

The demand for higher capacity and the limited availability of additional bandwidth call for more aggressive re-use of the spectrum via *cell densification*, which is accomplished by deploying small cells, or / and through the adoption of *uni-*

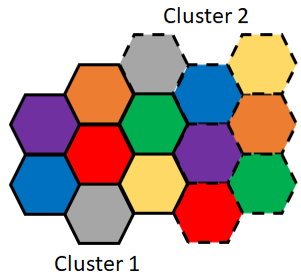


Figure 1.18 A frequency re-use pattern with re-use factor $N = 1/7$.

versal frequency re-use, where the available spectrum is used in each cell (i.e., the frequency re-use factor is unity). These strategies can lead to severe ICI, if no countermeasures are taken. ICI management techniques vary from *fractional frequency re-use* (*FFR*) in OFDMA systems, wherein the utilization of frequency re-use patterns is restricted to the cell boundaries to avoid the occurrence of ICI, to *coordinated multi-point* (*CoMP*) transmission schemes, which rely on the cooperation between neighboring BSs enable full frequency re-use and control the resulting ICI in the space domain [5, 32, 33].

1.9 5G Enabling Technologies

The *area throughput* of a cellular network, T_a, is defined as [27]:

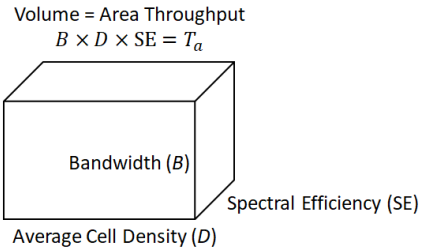
$$T_a \text{ [bit/s/km}^2\text{]} = B \text{ [Hz]} \times D \text{ [cells/km}^2\text{]} \times \text{SE [bit/s/Hz/cell]}, \tag{1.31}$$

where B is the *bandwidth*, D is the *average cell density*, and SE is the *spectral efficiency* per cell.

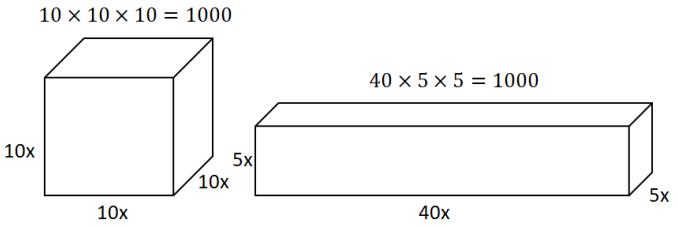
According to Eq. (1.31), we can visualize the area throughput as a rectangular box whose sides have length B, D, and SE and whose volume is T_a, as shown in Fig. 1.19a [27]. In Fig. 1.19b we see that we can reach the 5G capacity target by achieving different improvements of these factors [27].

An implication of Eq. (1.31) is that the network equipment vendors and MNOs have to rely on the synergy between the following strategies, in order to achieve the required 1000-fold increase of the capacity [1, 27, 34]:

1. Use of more bandwidth.
2. Densification of the network.
3. Improvement of the SE.



(a) Representation of the area throughput as the volume of a rectangular box with sides B, D, and SE.



(b) Different sets $\{B, D, \mathrm{SE}\}$ can achieve the 1000x capacity target.

Figure 1.19 Visualization of the area throughput and the 1000x capacity challenge [27].

1.9.1 Use of More Bandwidth

The main "vehicle" for achieving the 5G capacity goal is the spectral bandwidth. There exist several approaches regarding the extension of the usable spectrum.

Sub-6 GHz spectrum: The obvious way to improve the capacity is the "brute-force" approach of allocating additional bandwidth to 5G services. Traditionally, cellular networks have been utilizing sub-6 GHz spectrum. This is due to the characteristics of the wireless channel in these frequencies, which facilitate long-range communication as well as communication in scenarios where radio wave propagation is obstructed by large objects. However, the high demands for wireless connectivity on the one hand and the use of the spectrum on an exclusive-basis by the operators on the other resulted in the exhaustion of resources in this spectral segment.

Spectrum refarming: Spectrum refarming has been proposed as a workaround to the spectrum crunch issue. Nevertheless, this spectrum reallocation process is complex, time-consuming, and costly, since in most cases the spectrum under consideration has to be cleared and awarded to the operators via an auction scheme [35].

mmWave spectrum: The use of *millimeter-wave spectrum* (*mmWave*), which spans the frequency range 30–300 GHz, is a promising alternative, because of the sub-

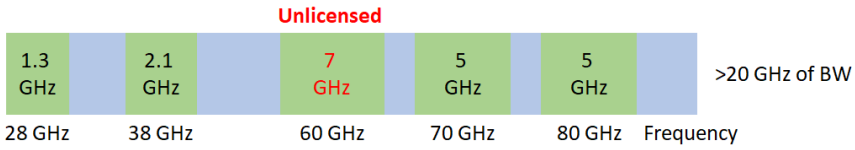


Figure 1.20 mmWave spectrum offers an enormous amount of bandwidth [40].

stantial amount of contiguous unexploited bandwidth in this spectral segment, as illustrated in Fig. 1.20 [16, 36–41].

Spectrum sharing: Another method for virtually acquiring additional bandwidth is the sharing of spectrum between two systems on a non-interfering basis. This concept dates back to the *cognitive radio* (*CR*) era. In general, an MNO can access the licensed spectrum of an incumbent either orthogonally or in a non-orthogonal manner [42–45] (see Fig. 1.21). *Orthogonal spectrum sharing* takes advantage of the allocated spectrum's severe underutilization [46]. In this spectrum usage model, the MNO relies on *spectrum sensing* or / and consults a *database* that stores information regarding the spectral activity of the incumbent(s) in the time-frequency-location domain, if such a registry has been made available by the NRA, to detect and subsequently exploit "spectrum holes" (i.e., temporarily idle channels). In *non-orthogonal* (or *underlay*) *spectrum sharing*, on the other hand, the players transmit concurrently over the same frequency band. In this spectrum management paradigm, the MNO makes use of techniques such as transmit BF and power control to restrict the power of the resulting CCI at the incumbent receivers below a predefined threshold.

1.9.2 Densification of the Network

A different approach for enhancing the capacity is the deployment of a large number of *small cells* to enable aggressive re-use of the available spectrum across the service area [47, 48]. Given the enormous capacity required by 5G networks, the use of *ultra-dense networks* (*UDN*), where the *inter-site distance* (*ISD*) is 100 m or less (about 10 m in indoor setups), seems a natural choice. Cell size shrinking and network densification is expected to provide the biggest capacity gains, in comparison to the other strategies [34].

1.9.3 Improvement of the Spectral Efficiency

Yet another direction is the employment of techniques that increase the SE per cell. This refers mainly to multi-cell MU-MIMO technologies.

Multi-user MIMO: Conventional (single-cell) MU-MIMO relies on *user scheduling*, *precoding*, and *power allocation* (*PA*) to mitigate or eliminate the intra-cell MUI, improve the receive SNR at the intended users, and increase the cell's SE [8, 49–51].

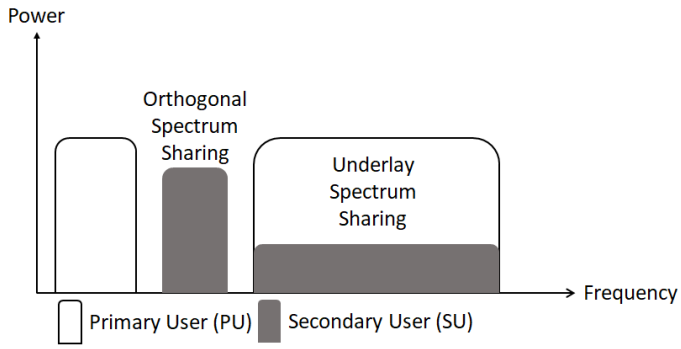


Figure 1.21 Simplified representation of the orthogonal and underlay spectrum sharing concepts.

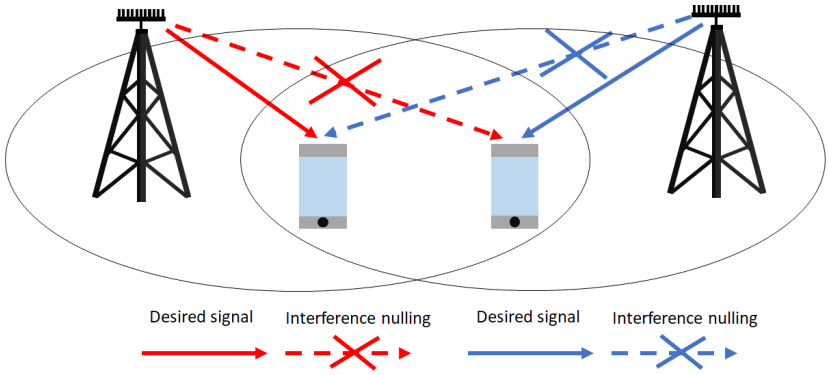
Precoding requires the availability of CSIT, which is obtained via feedback in FDD systems or via exploitation of the channel reciprocity in TDD systems. We distinguish between *instantaneous CSI*, which refers to knowledge of the instantaneous gain and direction of the channel, and *statistical CSI*, which implies knowledge of the distribution of the channel (or, often, just of its first and second order statistics, i.e., its mean value and its covariance). We further distinguish between *perfect CSI* and *imperfect CSI* caused by quantized feedback and feedback delays in FDD systems or by ICI between pilots (*pilot contamination*) in TDD systems. The latter is particularly important in massive MIMO systems.

Note that in MU-MIMO each BS determines selfishly its *resource allocation* (*RA*) policy and treats the ICI as additional noise. This approach is problematic in setups where the ICI is severe, since it leads to substantial performance degradation.

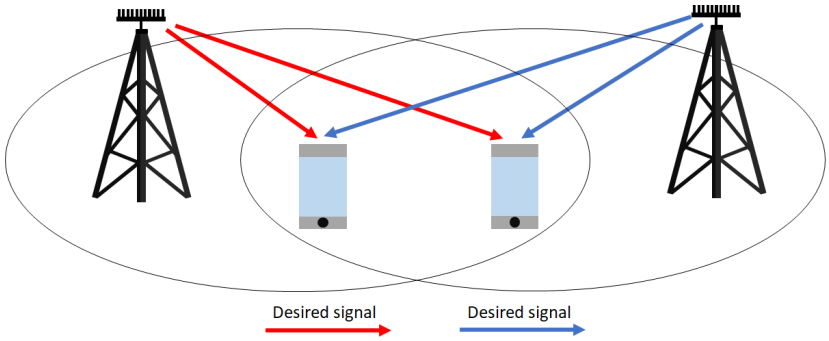
Coordinated / cooperative MU-MIMO: Coordinated / cooperative MU-MIMO constitutes a family of multi-cell MU-MIMO technologies wherein the BSs cooperate with each other to coordinate their RA strategies, so that the ICI is managed and the system-wide SE is increased [5, 32, 33, 50–52]. These technologies have been introduced in the Release 11 of *Long Term Evolution Advanced* (*LTE-A*) under the name *coordinated multi-point* (*CoMP*) [53] and is expected to constitute an integral component of 5G as well [34].

Coordinated MU-MIMO refers to partial cooperation strategies, where the BSs exchange CSI and possibly other control information over the mobile transport network to serve their users in a coordinated manner. Several coordination variants are defined, namely, *coordinated scheduling* (*CS*), *coordinated BF* (*CBF*) (also called *coordinated precoding*, *CP*), *coordinated PA* (*CPA*), and combinations thereof (see Fig. 1.22a).

Cooperative MU-MIMO, on the other hand, refers to full cooperation strategies, where the BSs share also user data. Under this paradigm, there exist two trans-



(a) Coordinated beamforming (CBF).



(b) Joint transmission (JT)

Figure 1.22 Examples of inter-cell coordination and cooperation [54].

mission options: either the scheduled users are served jointly by the cooperating BSs or they are served by a single BS that is selected dynamically out of them in each slot based on the channel conditions and the available resources. The former scheme is known as *joint transmission* (*JT*), whereas the latter one is called *dynamic cell selection* (*DCS*) (see Fig. 1.22b). JT and DCS are often referred to collectively as *network MIMO*. Notice that JT further improves the QoS of the cell-edge users. This is accomplished, though, at the expense of greater cooperation overhead, due to the sharing of user data, and strict synchronization requirements.

Note that typically inter-BS cooperation is restricted within clusters of BSs, to limit the cooperation and CSI acquisition overhead and enable the scaling of CoMP setups [5].

Massive MIMO: *Massive MIMO* (*mMIMO*) represents another multi-cell MU-MIMO technology, wherein each BS is equipped with an excessive number of antennas, in comparison to the number of active users (or to the number of transmitted data streams, in general, to include the case of multi-stream communica-

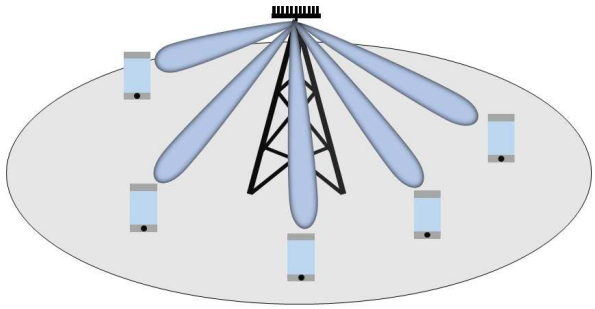


Figure 1.23 The excess of transmit antennas in mMIMO setups enables highly directional transmissions to multiple users. The small amount of residual interference is eliminated via simple linear precoding techniques.

tion with multi-antenna MSs) [27, 31, 55, 56] (see Fig. 1.23). The surplus of *spatial degrees-of-freedom* (*DoF*) provides high *array gain* and enable us to achieve high *multi-user spatial multiplexing* (*SM*) *gain* and efficient intra-cell and inter-cell CCI suppression without the need for inter-BS coordination / cooperation. The excess of transmit antennas leads also to *favorable propagation* (i.e., near-orthogonal user channels), which implies that simple linear precoding schemes are near-optimal from a sum-SE perspective. Furthermore, in this regime, small-scale fading vanishes (i.e., the instantaneous channel gain becomes almost equal to the average channel gain), thus simplifying channel estimation and power allocation. This phenomenon is known as *channel hardening*. mMIMO systems operate typically in TDD mode, where the number of UL pilots required for the estimation of the DL channels scales with the number of users and no feedback is needed.

1.10 Challenges, Misconceptions, and Opportunities

Although the above mentioned technologies enable us to improve the capacity towards the 1000x goal, each one of them faces a number of challenges that have to be addressed, for commercial implementation to take place or in order to exploit their maximum potential. In addition, there exist some common misconceptions regarding the features of these techniques. On the other hand, though, these methods offer also opportunities for effective synergies and enhancements.

1.10.1 mmWave Communication

The characteristics of mmWave channels (higher distance-dependent path loss and penetration losses, high probability of blockage from obstacles and human bodies, limited scattering and diffraction, higher noise power due to the larger bandwidth,

shorter coherence time / higher Doppler spread) limit the range and mobility support of mmWave communication. This fact answers why we can't rely solely on spectrum extension to address the 1000x capacity challenge: With the cellular networks collectively utilizing currently more than 1 GHz of bandwidth in the sub-6 GHz spectrum [27], we would have to use more than 1 THz of unexploited bandwidth. This is possible only via the application of mmWave and THz communication technologies, which, as mentioned previously, cannot meet the radio coverage and mobility management requirements of future cellular networks.

On the other hand, we should note that the mmWave spectrum is particularly appealing for use in small cells [1]. Moreover, the high array gain provided by mMIMO can compensate for the severe path loss of mmWave signals, while the short wavelength of mmWave frequencies facilitates the packing of hundreds of antennas in devices with small form-factor [1]. Also, the occurrence of blockage and the use of highly directive transmissions can be beneficial in the context of interference management in general [1] and, therefore, in spectrum sharing in particular [57, 58].

1.10.2 Spectrum Sharing

Despite its capacity enhancement capability, orthogonal spectrum sharing has been met with skepticism by the community so far, due to its lack of QoS provisioning. More specifically, the spectrum-sensing-based *opportunistic spectrum access* (*OSA*) approach suffers from *misdetection errors* and *false alarm errors*, which refer to the situation where an occupied channel is declared as idle or an idle channel is declared as occupied, respectively. The former type of errors leads to interference, while the latter type reduces the potential capacity gain due to missed shared spectrum access opportunities. The inability of stand-alone single-antenna sensors to detect reliably the activity of incumbents (as in the well-known *hidden node problem*), mainly due to shadowing and multipath propagation effects, led to the development of various alternatives. The use of multiple antennas at the sensors represents one example [59–63]. The array gain and the diversity gain of the multi-antenna sensing nodes improve the sensing reliability in low-power / high-interference scenarios and in multipath propagation environments, respectively. A more popular approach is the application of *collaborative spectrum sensing*, which exploits the spatial diversity in the observations of sensing nodes located at different places to obtain more reliable spectrum sensing outcomes through centralized or distributive cooperation at the expense of the additional cooperation overhead [64–67]. However, while these methods improve the performance of spectrum sensing, they are still subject to misdetection errors and false alarm errors.

The lesson learned from previous experience is that database-assisted spectrum sharing is the only viable solution for efficient orthogonal access on shared spec-

trum. The first system that put forth this concept was *TV White Spaces* (*TVWS*), wherein unlicensed *secondary users* (*SU*) exploit the spectral gaps between the TV broadcasting frequency bands [68–70]. In particular, in TVWS SUs send to a database their current location and the accuracy of this measurement. Then, the TVWS database, which stores the location and other relevant transmission characteristics of the *primary users* (*PU*) (i.e., the TV broadcasting stations), such as their transmission power and antenna height, calculates a safety zone based on a radio wave propagation model and replies by sending a list of available channels and the allowed transmission power level.

TVWS faces many challenges, which are mainly related to its unlicensed secondary access nature. For instance, the SUs have to query frequently the database, therefore increasing the energy consumption and traffic load. Furthermore, the location of fast moving SUs may change before the database responds with the list of available channels, thus triggering again the initiation of this query-reply procedure. Also, the opportunistic use of the spectrum by the SUs does not allow for the provision of any QoS guarantees. In fact, TVWS "overprotects" the PUs, thus leading to unavailability of channels in dense urban environments. Similarly, often the calculation of the safety zones is inaccurate, thus resulting in harmful interference. Moreover, there is no mechanism for coordinating the transmissions between SUs (e.g., TVWS devices, wireless microphones and other *program making and special equipment* (*PMSE*) nodes, etc.). Finally, there are no financial incentives for the PUs. This fact, in conjunction with the inability to ensure the provision of prescribed QoS levels, results in lack of a clear business model. These characteristics, along with the advent of the digital dividend which reduced the available TVWS resources, prevented the commercial success of this technology.

Licensed Shared Access (*LSA*) represents a recently introduced frequency-agnostic database-assisted orthogonal spectrum sharing framework that addresses these issues [9, 13, 71–76]. LSA is an extension of the *Authorized Shared Access* (*ASA*) concept, which was proposed by NSN (now Nokia Bell Labs) and Qualcomm [77, 78]. Both paradigms refer to individual licensing for exclusive access to shared spectrum. In contrast to ASA, though, LSA includes also other types of licensees besides MNOs. LSA is promoted in Europe by CEPT and ETSI as a complementary licensing regime to licensed spectrum access for the 2.3–2.4 GHz band, which is used mainly by PMSEs (but also by military radars, amateur radio systems, etc.). In LSA, the incumbents sign long-term spectrum sharing agreements with the LSA licensees that state the frequency range of the spectrum under sharing consideration, the geographic areas and the time intervals over which the LSA licensees can access this shared spectrum, the spectrum usage rules, and the conditions that result in termination of spectrum sharing. These agreements include the definition of the *exclusion zones*, where the LSA licensees are not allowed to transmit; the *restriction zones*, where the transmissions of the LSA licensees are subject to certain constraints; and the *protection zones*, where the incumbent receivers should not

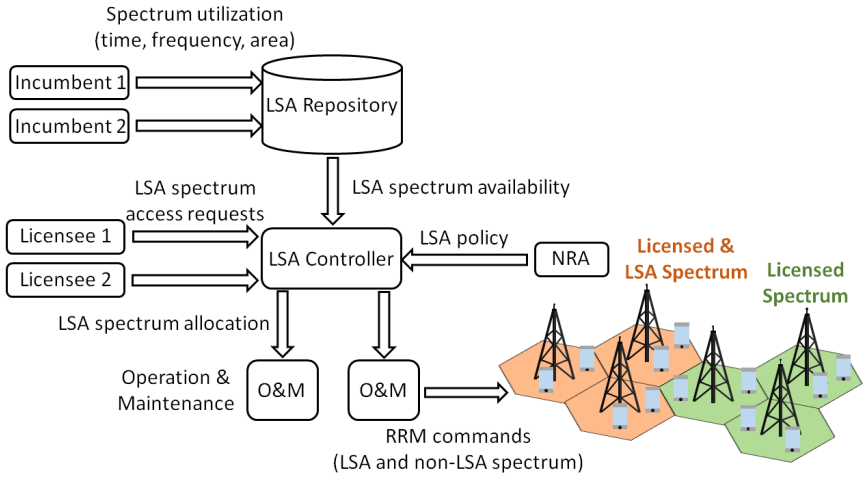


Figure 1.24 Key players and components of LSA.

be subject to harmful interference. Information regarding the incumbents' spectrum utilization in the time, frequency, and space domains (i.e., information on the LSA spectrum availability) is stored in the *LSA repository*, which is maintained by the NRA or a trusted third party. The *LSA controller*, on the other hand, accepts requests for access on the shared spectrum by the LSA licensees, computes the permitted LSA spectrum based on the LSA spectrum availability and the spectrum usage rules, and either allocates the requested spectrum to the corresponding operator or rejects the request. The LSA controller ensures interference-free operation and predictable QoS for both incumbents (tier-1 players) and LSA licensees (tier-2 players). The operations and maintenance (O& M) entity of each operator's network is responsible for the internal management of the allocated spectrum (considering both LSA and non-LSA spectrum). The key components of the LSA architecture, their functionality, and the LSA players are illustrated in Fig. 1.24.

Note that this subletting or temporary leasing of spectrum to LSA licensees defines a business model that gives rise to a new revenue flow for the incumbents and constitutes a fast and cost-effective way for MNOs to enhance their capacity at specific locations and times by aggregating 5G carriers with LSA carriers. Notice also that, in principle, the incumbents could be MNOs sharing their spectrum with smaller local service providers or with *mobile virtual network operators* (*MVNO*), although this business case is not defined in the standard.

Of course, LSA presents also some drawbacks. The operation of LSA is based on long-term static or semi-dynamic contracts between the incumbents and the LSA licensees. This conservative approach is highly inefficient, in the sense that it does not exploit fully the potential capacity gains of spectrum sharing. To this

end, *dynamic LSA* [79] and *evolved LSA* (*eLSA*) [80] have been proposed as enhancements of the conventional LSA paradigm. The former method complements the LSA architecture with a spectrum sensing network to make possible the dynamic detection of additional shared spectrum access opportunities, whereas the latter one combines spectrum sensing with short-term LSA licenses, local LSA spectrum allocations, and co-primary sharing to enable efficient and flexible allocation of the LSA spectrum and facilitate the introduction of novel services and business models.

Secondary Access System (*SAS*) is a similar to LSA spectrum sharing paradigm that has been initiated by the FCC for shared use of the 3.55–3.7 GHz CBRS band [13, 81]. It has, however, some important differences from LSA. More specifically, the main type of incumbents in the case of SAS is military services, which cannot provide any information to a database a priori. Therefore, SAS defines exclusion and protection zones, but relies ultimately on spectrum sensing to monitor spectrum utilization and perform spectrum allocation. The SUs are allowed to transmit within the protection zones, if this is dictated by the spectrum sensing outcome. Furthermore, SAS defines a third tier of unlicensed users, which can utilize opportunistically vacant spectrum, thus enjoying "best-effort" access. Finally, SAS requires the application of interference mitigation techniques, since: (i) the geographic areas over which the SUs can access the shared spectrum in dense urban environments is typically small, and (ii) the transmissions of the unlicensed users should be coordinated, to protect the tier-2 users. In essence, SAS enables more efficient / dynamic use of the spectrum than LSA, at the cost of less certain QoS.

Turning next our attention to underlay spectrum sharing, we notice that this method has been considered mainly in low-power / short-range communication scenarios, due to the fact that the *interference power threshold* (*IPT*) of the incumbents is usually very strict (i.e., within tolerable noise levels). However, technologies such as CoMP and mMIMO could offer high SE gains whilst protecting the incumbents from harmful interference in underlay spectrum sharing setups, thanks to their coordinated interference management and RA capabilities or to their highly directional transmissions and large number of spatial DoF, respectively [82, 83]. CoMP / mMIMO enabled underlay spectrum sharing could be also efficiently combined with small cells, where short-range communication takes place.

We should note that there is a recent interest in underlay spectrum sharing at mmWave frequencies, as a means to reduce the license costs and utilize the spectral resources more efficiently [57, 58]. This trend has been motivated by the spatial focusing of the transmissions in mmWave mMIMO systems via the shaping and steering of narrow beams, which reduces the inter-system CCI.

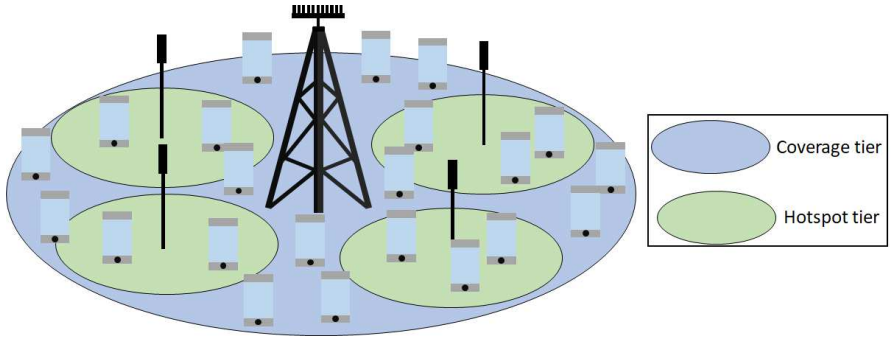


Figure 1.25 5G multi-tier multi-band network structure. Macro-cells operating at sub-6 GHz frequencies handle radio coverage and mobility management tasks, while small cells utilizing sub-6 GHz and mmWave carriers act as hotspots for capacity enhancement [27].

1.10.3 Network Densification

A dense network of small cells cannot meet alone the radio coverage and mobility support demands of contemporary and future cellular networks. Therefore, the deployment of UDNs makes sense only in the context of *heterogeneous networks* (*HetNet*), as shown in Fig. 1.25. We note that a tier of densely installed small cells that act as hotspots to provide local capacity enhancement overlays a macro-cell tier, which is responsible for handling the aforementioned tasks [27, 84]. A common implementation of this paradigm is a *control plane / data plane separation architecture* (*CDSA*) [85]. Also, cell densification is limited by network deployment and management costs and is hampered by the occurrence of ICI [27, 86]. However, we should mention that the *cloud radio access network* (*RAN*) architecture [87] reduces the *total cost of ownership* (*TCO*) and facilitates the utilization of ICI mitigation techniques, such as CoMP, thus enabling us to push further the limits of network densification.

1.10.4 Coordinated Multi-Point

One of the main implementation challenges of CoMP is that it places a heavy capacity and latency burden on the mobile backhaul network due to inter-BS cooperation [5, 51, 54]. *Cell clustering* schemes and the cloud RAN architecture facilitate multi-cell coordination. The former approach restricts inter-cell cooperation / coordination within clusters of cells. In cloud RAN setups, on the other hand, *virtual baseband units*(*vBUU*) are gathered at a *central office* (*CO*) and communicate with the *remote radio units* (*RRUs*), which are located at the cell sites, via an optical transport network. This network segment that connects the vBBUs to the RRUs is called the *mobile fronthaul* (*MFH*), as opposed to the *mobile backhaul* (*MBH*) that connects the RAN to the *core network* (*CN*).

Yet, the cooperation overhead might be prohibitively high, especially in the case of CoMP-JT, which entails user data and global CSI sharing [5, 51, 54]. *Mobile edge caching* has been proposed as a workaround to this problem. In this paradigm, servers that have been installed at the network edge (e.g., at the cell sites) store frequently requested content to serve future user requests locally, thus reducing the latency and the network traffic [88–90]. The redundancy of the stored content across the cache servers creates *JT opportunities* while completely *eliminating the need for user data exchanges* [91].

Coherent CoMP-JT presents also stringent synchronization requirements [5, 51, 54]. Furthermore, the sharing of global CSI in each cluster complicates channel estimation and puts high demands on the CSI feedback links [5, 51, 54]. In addition, although centralized RA typically improves the performance in comparison to distributed RA, the corresponding algorithms might be infeasible due to high computational demands and delays [5].

Another challenge is the need for a clustering algorithm that demonstrates a favorable performance vs. complexity trade-off. Cell clustering in CoMP setups is performed in either a *network-centric* (*NC*) or a *user-centric* (*UC*) manner [54]. NC clustering refers to the formation of predetermined disjoint cooperation clusters based on a network perspective (e.g., cell's adjacency) in either a *static* or a *semi-dynamic* manner. Networks with static clusters provide poor overall SE when the user distribution is heterogeneous [32] and suffer from *out-of-cluster* (*OOC*) interference, which further degrades the sum-SE [92]. Using frequency planning on a cooperation cluster level [93] or on cluster-edge level [93] or changing over time the non-overlapping clusters [32, 94] improves the performance. Nevertheless, these strategies do not address the aforementioned issues, which are related inherently with the NC formation of cooperation clusters that transforms the ICI management task into an OOC interference management task (i.e., under NC clustering, the cooperation clusters play essentially the role of large cells, from an OOC interference management perspective). In UC clustering, on the other hand, the BSs form dynamically different (possibly overlapping) cooperation clusters for different users from a user perspective, i.e., based on parameters such as the location of the users, their QoS requirements, etc. [92, 95–99]. Thus, UC clustering solves the problems associated with the NC approach by "blurring" the concepts of cell and user-cell association, at the cost of extra computational load[15].

Finally, it is well known that the performance of CoMP improves in general with the number of BS antennas [8, 100], since the additional spatial DoF enable us to increase either the multi-user SM gain or the array gain for a given SM gain [100]. Hybrid analog-digital TRXs that make use of *load-controlled parasitic antenna arrays* (*LC-PAA*) enhance, in principle, the performance for a given number of RF chains. Their operation is based on the use of closely-spaced active and passive antennas, with the latter being terminated to adjustable loads, and the ex-

[15]The concept of BF makes less relevant the definition of the cells as well.

ploitation of the strong electromagnetic coupling among them. The *mutual coupling* enables the passive antennas to radiate, while the adjustment of the loads allows us to control dynamically the shape of the radiation pattern or to perform channel-dependent precoding [100]. The adaptive computation of the loads that determine the currents on the passive antennas, though, is a challenging task. Furthermore, the precoded signals may correspond to infeasible load values.

1.10.5 Massive MIMO

The large number of BS antennas in mMIMO implementations necessitates the use of efficient antenna array designs and TRX architectures [1]. The difficulty in mMIMO implementations arises from the large number of RF chains—especially in the case of mmWave mMIMO. This is because in fully digital TRXs, each antenna is connected to an RF module and a *digital-to-analog / analog-to-digital converter* (*DAC / ADC*). The cost and and power consumption of such a mmWave mMIMO implementation is prohibitively high with current technology.

A solution to this problem is the use of *hybrid analog-digital TRXs*, to limit the number of RF chains and ADCs. In such TRXs, part of the precoding / combining operations is performed in the digital baseband and part of them is performed in the analog domain via a network of phase shifters or switches [101]. Hence, there is a trade-off between performance and cost / power consumption, which is related with the limitations of analog processing.

More specifically, the phase shifters typically support only quantized phases (i.e., discrete phase shifting) and impose a constant modulus constraint in the design of the analog precoding / combining matrix which makes the SR maximization problem non-convex. Therefore, one has to rely on approximations that perform close to the unconstrained fully-digital solution—e.g., by exploiting the sparsity or low rank of the mmWave mMIMO channel matrix[16] [102, 103]. Note that the analog precoder is constant also across all subcarriers, thus further complicating the TRX design of multicarrier systems [104].

On the positive side, the highly directional transmissions of mMIMO systems facilitate interference management and underlay spectrum sharing, as we have already mentioned [83].

1.11 Resource Allocation

Resource allocation (RA) refers to a set of techniques whose goal is to optimize the performance at network level and provide fairness and QoS guarantees at user level, according to predefined metrics [51].

[16]Channel sparsity refers to the fact that there exist only a few paths between the TX and the RX, due to the limited number of scattering clusters. The channel is characterized by the *angle of departure* (*AoD*), *angle of arrival* (*AoA*), and *gain* for each of these paths.

A *resource allocation policy* consists of [51]:

- a multiple access scheme and a user scheduling algorithm, to distribute the resources among the users according to the performance criterion and the QoS requirements;
- a signaling strategy, to enable transmissions to multiple users based on the applied multiple access and user scheduling techniques; and
- a rate adaptation and power allocation (PA) method, to ensure compliance with the QoS demands.

Multiple access schemes are divided into orthogonal and non-orthogonal, as we have seen in Section 1.8. The former avoid the occurrence of co-channel interference by allowing a BS to serve multiple users either on different frequencies or at different times. Thus, their multi-user multiplexing gain is limited by the number of available resources (sub-carriers and timeslots) [51]. Non-orthogonal schemes, on the other hand, introduce an additional signaling dimension to enable communication with multiple users on a single time-frequency resource, thus increasing the SE. The management of the resulting CCI relies on advanced transmission techniques and signal processing operations performed at the TX (pre-processing) or / and the RX (post-processing).

MU-MIMO constitutes a characteristic example. This technology essentially takes advantage of the spatial DoF offered by the multiple antennas at the BS to spatially multiplex independent data streams that are destined to different users. Intra-cell MUI mitigation is based on channel-dependent precoding. The optimal precoding strategy is *dirty paper coding* (*DPC*), which exploits the non-causal knowledge of intra-cell MUI to eliminate it prior to transmission [8]. However, the successive encoding and decoding operations of this non-linear pre-processing method make it impractical. Among the suboptimal non-linear and linear alternatives, the latter provide a better trade-off between computational complexity and performance. The goal of linear precoding schemes is to balance the increase of the receive SNR at the intended users and the reduction of the interference at the non-intended users [5]. Nevertheless, for most problems of interest, the computational complexity of the optimal linear precoding strategy is prohibitively high [51]. Therefore, we commonly rely in practice on simple heuristics.

The spatial multi-user multiplexing gain of MU-MIMO setups that utilize fully digital transceivers is determined by the number of transmit antennas. If the transmission of a larger number of data streams is required, user selection should be applied, according to network performance, user fairness, and QoS criteria. Power allocation can further improve the performance at network or / and user level.

Resource allocation is a particularly challenging task due to [5, 51]:

- the multi-dimensional pool of resources (subcarriers, timeslots, powers, antennas, users);

- the inherent trade-off between SR maximization (which is often the network performance criterion) and user fairness;
- the fact that the resource allocation components are intertwined;
- the inter-user coupling through interference, which typically turns the optimization tasks into non-convex problems;
- the combinatorial nature of user selection; and
- the randomness and time-variability of the wireless channel.

Coordinated multi-cell resource allocation and dynamic user-centric clustering further complicate the resource allocation (e.g., the BSs become a resource as well). Note that dynamic cell clustering and user selection are tightly coupled.

1.12 Motivation and Goals

Based on the discussion so far, we make the following observations:

- The sub-6 GHz spectrum plays an important role in the 5G ecosystem, since it provides the means to meet the radio coverage and mobility support requirements of the 5G standard [27].
- Given the scarcity of resources in this region of the spectrum [9], the operators should exploit the full potential of the applicable capacity enhancement strategies, in order to address the 1000x challenge. These include [9, 27]: (i) the use of transmission methods that enable more efficient utilization of the available spectrum in each cell; (ii) the adoption of technologies that mitigate the ICI, thus increasing the overall SE and facilitating cell densification; and (iii) the application of spectrum sharing, to enable access to previously reserved spectrum, thus increasing the effective system bandwidth. CoMP and mMIMO fall under classes (i) and (ii), whereas LSA falls under class (iii).
- LSA ensures interference-free operation for both players by enforcing orthogonal access to the shared spectrum based on database-assisted spectral activity detection and a long-term spectrum usage agreement [72, 79, 80]. However, the stringent capacity requirements of 5G dictate the need for more aggressive spectrum sharing.

Under this context, the main motivation of our study is summarized in the following statement:

Multi-cell MU-MIMO technologies present advanced interference management and resource allocation capabilities, thanks to inter-cell cooperation in the case of CoMP and

to the excessive number of spatial DoF in the case of mMIMO. Therefore, a CoMP- or mMIMO-enabled underlay spectrum sharing paradigm has the potential to achieve substantial sum-SE gains while ensuring the provision of QoS guarantees to the end users.

The study of this combination of modern spatial-domain-based interference management and resource allocation techniques with underlay spectrum sharing constitutes the main topic of this dissertation. *We believe that the features of the proposed spectrum sharing paradigm can enable its incorporation into the LSA framework, thus leading to a new generation of spectrum sharing techniques that combine orthogonal and non-orthogonal spectrum sharing with QoS guarantees to further extend the usable spectrum*.

Another objective is the application of this concept at mmWave frequencies, where there have been conducted already some studies regarding the efficient inter-operator sharing of the spectral resources. mmWave spectrum sharing has been primarily motivated by the high directivity of the transmissions and the high probability of blockage in such high frequencies, which facilitate interference management [1, 57, 58], as well as by the interest of fixed and satellite systems to use the mmWave spectrum in the future.

Let us take a look now at the individual objectives of this research work.

- *The bulk of the literature focuses on the conjunction of the underlay spectrum sharing paradigm with legacy technologies*, such as uncoordinated MU-MIMO.
- Furthermore, the majority of the few relevant studies *does not consider the QoS requirements of the end users*.
- Also, these studies on sum-rate (SR) maximization typically neglect the user scheduling procedure (i.e., they consider an arbitrary set of active users).
- In addition, *the application of standard linear precoding schemes in CoMP-enabled underlay spectrum sharing is considered only in some special cases* [5]. This *pragmatic* strategy, where well-known robust transmission methods are utilized, reduces the implementation complexity and could accelerate the adoption of this spectrum sharing solution by commercial deployments.
- Moreover, *CoMP-JT is rarely used in practice*, since it imposes a heavy burden on the mobile transport network in terms of throughput and latency requirements. *Cache-aided CoMP-JT* has been proposed as a workaround to this problem. However, this technology has not been studied under the underlay spectrum sharing context. Besides, most studies consider uncoordinated transmissions when cache-aided CoMP-JT cannot take place [105, 106]. This approach is highly suboptimal from a sum-SE maximization and interference management perspective and does not suit the underlay spectrum sharing context under consideration.

- The performance of CoMP transmission techniques is directly related with the number of antennas at each BS. We can improve the performance by using *LC-PAAs*, as we have mentioned. *The use of such hybrid analog-digital TRXs, though, has not been studied under underlay spectrum sharing—or in the context of CoMP transmissions, for that matter*.
- Finally, *mmWave mMIMO has not been studied yet as an enabler of underlay spectrum sharing*, to the best of our knowledge.

Our goal is to fill these gaps in the literature.

1.13 List of Contributions

Let us list the contributions of this work:

- We derive *low-complexity coordinated precoding, power allocation, and user scheduling schemes for QoS-aware and QoS-agnostic SR maximization* in underlay spectrum sharing setups. *Standard linear precoding schemes or simple variations of them, such as zero-forcing (ZF) and projected ZF precoding, are utilized*. Different precoding schemes are proposed for different interference power threshold regimes. The developed coordinated user scheduling schemes are based on search space reduction or on the exploitation of the cross-correlation between the user channels and the inter-system interference channels. *Greedy implementations* are also proposed.
- A *cache-aided CoMP-JT scheme based on a coordinated content caching with redundancy enhancement* (*C3RE*) *method and efficient frequency-/recency-based caching schemes, namely, score-gated least recently used* (*SG-LRU*) *and score-gated clock* (*SG-C*), is proposed.
- The utilization of *load-controlled parasitic antenna arrays* (*LC-PAA*) is considered. *Coordinated hybrid precoding* is studied under this context. Also, a *beam selection and precoding* (*BSP*) technique is proposed, which decouples analog beamforming from digital precoding, as a workaround to the challenges of load computation for arbitrary channel-dependent precoding with LC-PAAs.
- The use of *coordinated symbol-level precoding* is studied, which improves the performance in the low SNR regime.
- Finally, *hybrid processing for mMIMO and mmWave mMIMO links* is studied and an efficient *hybrid precoding / combining algorithm based on stochastic approximation with Gaussian smoothing* (*HPSAGS*) is derived.
- The performance of the proposed techniques is evaluated for *a variety of primary system (PS) setups and operating parameters* via *numerical simulations at*

both cluster (or cell) and system level. System-level simulations are based on the *non-line-of-sight (NLOS) macro-cellular 3GPP model for carrier frequency of 2 GHz* [107]. Moreover, a *dynamic cell clustering* (*DCC*) scheme is described and applied in system-level simulations.

1.14 Structure of the Dissertation

The structure of the dissertation is as follows: In Chapter 2 we study the application of coordinated resource allocation in underlay spectrum sharing setups. In Chapter 3 we focus on cache-aided joint transmission. Chapter 4 considers coordinated hybrid codeword-level and symbol-level precoding. Chapter 5 deals with hybrid precoding for massive MIMO setups, with focus on mmWave frequencies (although the proposed technique can be applied at sub-6 GHz as well). Finally, in Chapter 6 we provide a summary and present the conclusions of this work.

Chapter 2

Spectrum Sharing I: Coordinated Resource Allocation

2.1 Introduction

The mobile network operators rely on the synergy between a number of strategies to meet the extreme capacity requirements of the 5G standard [108]. Examples include the utilization of mm-wave spectrum, where there is an abudance of unexploited spectral bandwidth, and the dense deployment of small cells, which enables more frequent re-use of the available spectrum across the service area [9, 27]. Nevertheless, the former approach is suitable only to short-range applications due to the high distance-dependent path loss in these frequencies, whereas the latter one is hampered by the occurrence of inter-cell interference (ICI) [9, 27]. These limitations indicate the key role of the highly congested sub-6 GHz spectrum in 5G and highlight the importance of techniques that enable access to additional spectrum, alleviate the ICI, or increase the spectral efficiency (SE) [9, 27].

Coordinated precoding / scheduling (CP / CS) [5, 51] and spectrum sharing [42, 44] constitute characteristic examples of such technologies. In CP / CS neighboring base stations (BS) cooperate with each other to coordinate their resource allocation (RA) policies, so that both the intra-cell multi-user interference (MUI) and the ICI are mitigated and the overall SE is increased. CP / CS facilitates network densification and improves the quality-of-service (QoS) of the cell-edge users. Inter-cell cooperation is typically restricted within clusters of BSs, to limit the signaling and channel state information (CSI) acquisition overhead [5]. In spectrum sharing, on the other hand, the operator access the licensed spectrum of an

incumbent either orthogonally in time or / and frequency or in parallel with the incumbent. In the former case, spectrum sharing is based on spectrum sensing or database-assisted incumbent's activity detection, whereas in the latter one it relies on the utilization of techniques that maintain the resulting inter-system co-channel interference (CCI) at the incumbent users below a predefined threshold.

The community has been reluctant to adopt spectrum sharing, despite its capacity enhancement nature, because of its lack of QoS provisioning. Licensed shared access (LSA) addresses this issue by enforcing orthogonal access to the shared spectrum based on database-assisted spectral activity detection and a spectrum usage agreement [72, 75].

2.1.1 Motivation

The stringent capacity requirements of 5G call for more aggressive spectrum sharing than LSA. The combination of non-orthogonal (or underlay) spectrum sharing with CP / CS promises substantial SE gains and the provision of QoS guarantees, thanks to coordinated RA and interference management. Therefore, it is worth to study the incorporation of CP / CS enabled underlay spectrum sharing in the LSA framework, as a means to further extend the usable spectrum.

However, one should first revisit the underlay spectrum sharing problem under the aforementioned context. The few relevant research works typically either consider an arbitrary set of active users (i.e., they neglect user selection) or / and deal with sum-rate (SR) maximization problems, which result in QoS-agnostic RA strategies. Also, the majority of the studies that treat CP / CS as an enabler of underlay spectrum sharing do not consider the application of well-known robust linear precoding techniques, as a pragmatic approach that could accelerate the adoption of this spectrum sharing solution by commercial deployments.

2.1.2 Related Work

The maximization of the DL SR in an underlay spectrum sharing setup has been studied extensively in the past for use cases where the secondary system (SS) is either a multiple-input multiple-output (MIMO) link or a MIMO broadcast channel (MIMO BC) and is collocated with one or more single- or multi-antenna primary receivers (e.g., see [109–112]). Popular approaches include: (i) the use of singular value decomposition (SVD) based predcoding/combining to cancel the inter-stream interference; and (ii) the application of zero-forcing (ZF) based precoding to eliminate the inter-user CCI within the SS or the inter-system CCI at the primary receivers when their interference power threshold (IPT) is null. Since these strategies remove the coupled interference, they convert the optimization problem into a convex PA task whose solution is either the standard water-filling (WF) PA scheme, when the inter-system CCI has been eliminated, or an interference-constrained variant of it that converges to the standard WF-PA for sufficiently

large IPTs, otherwise. This solution is commonly obtained via an iterative algorithm that makes use of standard optimization tools.

In contrast to the plethora of such research works, there are only a few corresponding studies that consider the maximization of a secondary coordinated multi-user MIMO network's DL SR or weighted SR (WSR), where the weights represent user priorities. For instance, in [5] the authors derive the optimal ZF-based precoding scheme for a scenario where the primary receivers do not tolerate any interference. In [113] a MIMO interference channel (IC) setup with single-antenna mobile stations (MS) that is collocated with a number of single-antenna primary receivers is considered. The authors determine the beamforming (BF) scheme that maximizes the WSR via an iterative algorithm that is based on the subgradient method, under the assumption that all secondary BSs have full knowledge about the channels of the secondary network as well as about the cross channels that link them to the primary receivers. Finally, the authors in [82] consider a secondary MIMO IC with multi-antenna MSs that coexists with a number of primary single-input single-output (SISO) links. They apply a semi-definite relaxation to the problem, derive a first order linear approximation of the SR via Taylor series expansion, and determine the optimal transmit BF and receive combining schemes for the derived convex problem by utilizing a computationally efficient iterative algorithm that successively optimizes the transmit beamformers for given receive combiners and vice-versa. They study also the problem of fairness optimization under a similar framework.

2.1.3 Contributions

In this work, we aspire to fill the aforementioned gaps in the literature by deriving coordinated RA strategies that maximize the DL SR under minimum rate constraints (MRC) and consist of standard linear CP schemes, simple coordinated PA methods, and efficient heuristic CS algorithms.

More specifically, in this chapter we consider a secondary MIMO interference broadcast channel (MIMO IBC) with single-antenna MSs and coordination among the BSs. The network adopts universal frequency re-use, meaning that the same frequency channels are utilized in all cells. The multi-antenna BSs utilize CP to serve their own users on a single time-frequency resource in a coordinated manner.

We consider three primary system (PS) setups:

1. A SISO link.
2. A MIMO link.
3. A MIMO BC.

Moreover, we focus on two scenarios:

1. The SS (i.e., the cellular network) operates under per-BS sum-power constraints (SPC) and an interference power constraint (IPC) associated with each primary receiver.

2. The transmissions of the BSs are subject also to per-user MRCs, which represent QoS requirements.

Initially, we consider an arbitrary set of active users and study the SR maximization problem for a simple setup where all cells belong to the same cooperation cluster. We are interested in the study of this system for the case where standard coordinated linear precoding schemes or simple variations of them are applied at the cellular network. We consider two approaches:

1. Coordinated ZF (C-ZF) precoding is utilized in a spectrum-sharing-agnostic manner, i.e., the inter-system CCI at the primary receiver(s) is ignored and the inter-system CCI at the MSs is treated as additional noise. In this case, the intra-system CCI, i.e., the intra-cell MUI and the ICI, is eliminated. We derive coordinated PA schemes to protect the PS from harmful interference and maximize the SR of the SS for the given transmission constraints and precoders. Depending on the scenario under study (i.e., depending on whether we consider SR maximization under MRCs or not), we develop both QoS-aware and QoS-agnostic coordinated PA schemes. These PA strategies can be applied heuristically also to other linear precoding schemes, such as coordinated regularized ZF (C-RZF) precoding.

2. Coordinated projected ZF (P-ZF) precoding is utilized, in order to cancel the inter-system CCI at the primary receiver(s).

Next, we study the CS or user selection problem. This problem is combinatorial. The optimal solution entails exhaustive search over a multi-user multi-cell search space. The computational complexity of this method is prohibitive. Therefore, we focus on suboptimal alternatives. Two such schemes are proposed:

1. Reduced search space (RSS) based user selection.

2. Inter-system cross-correlation aware user selection.

Finally, we consider a multi-cluster setup, where inter-cell coordination is restricted on a per-cluster basis. Cluster formation is dynamic and is determined by the achieved sum-SE. The parameters of the simulation are based on the non-line-of-sight (NLOS) macro-cell 3GPP model for 2 GHz carriers [107].

The performance of the proposed resource allocation schemes is evaluated via an extensive set of numerical simulations for various parameters. These simulations reveal that the interference-constrained PA schemes or projected ZF precoding enable efficient use of the spectrum and QoS provisioning when the IPT is

relaxed or hard, respectively. They indicate also that RSS coordinated user scheduling performs close to the optimal user selection approach.

Notice that we follow a step-by-step approach: First, we consider a single cluster setup with an arbitrary set of active users and present CP and coordinated PA solutions to the SR maximization problem with or without MRCs. Then, we introduce CS techniques. Finally, we consider a multi-cluster setup and we incorporate large-scale fading effects and cell clustering. The reason for this approach is because it enables us to evaluate the effect of each component on the performance of the system (i.e., coordinated precoding and power allocation; coordinated user scheduling; large-scale propagation phenomena; and cell clustering).

2.1.4 Organization and Mathematical Notation

The remainder of this chapter is organized as follows: In Section 2.2 is described the system setup and are introduced the main assumptions that we consider in our study. Section 2.3 presents the signal models for the various setups of interest. In Section 2.4 are formulated the problems under study and are presented the proposed CP and coordinated PA schemes, along with the implementation algorithms of these PA methods. Section 2.5 presents the CS problem and strategies. The results of the numerical simulations are shown in Section 2.6. Finally, the summary and conclusions of this work are presented in Section 2.7, while the proofs of the mathematical Theorems are given in the Appendix.

Mathematical Notation: $\mathbb{C}$ and $\mathbb{R}$ denote the set of complex and real numbers, respectively, while $\mathbb{R}_+$ denotes the set of non-negative reals. $a \in \mathbb{C}$ is a complex-valued scalar. $\mathbf{a} \in \mathbb{C}^n$ represents a column-wise n-dimensional vector with complex-valued elements. $|a|$ denotes the magnitude (absolute value) of a complex-valued (real-valued) scalar a. $\mathbf{A} \in \mathbb{C}^{n \times m}$ represents a $n \times m$ matrix $\mathbf{A}$ with complex-valued entries. $(\mathbf{a})_i = a_i$ denotes the i-th element of $\mathbf{a}$ and $(\mathbf{A})_{ij} = a_{ij}$ represents the (i,j)-th entry of $\mathbf{A}$. $(\mathbf{A})_{i*}$ and $(\mathbf{A})_{*j}$ denote the i-th row and j-th column, respectively, of $\mathbf{A}$. $\|\mathbf{a}\|$ and $\|\mathbf{A}\|$ stands for the Euclidean norm of $\mathbf{a}$ and $\mathbf{A}$, respectively. $\mathbf{A} = \mathrm{diag}\,(a_1, \ldots, a_n)$ is a diagonal matrix whose on-diagonal elements are $a_{ii} = a_i$, $i = 1, \ldots, n$. $\mathbf{A}^T$, $\mathbf{A}^\dagger$, and $\mathbf{A}^\# := \mathbf{A}^\dagger \left(\mathbf{A}\mathbf{A}^\dagger\right)^{-1}$ denote the transpose, Hermitian transpose, and Moore-Penrose pseudo-inverse, respectively, of $\mathbf{A}$. $\mathbf{I}_n$ and $\mathbf{0}_n$ denote the $n \times n$ identity matrix and the n-dimensional null vector, respectively. $a \sim \mathcal{N}\left(0, \sigma^2\right)$ and $a \sim \mathcal{CN}\left(0, \sigma^2\right)$ represents a real-valued and a circularly-symmetric complex-valued, respectively, Gaussian variable a with zero mean and variance σ^2, while $\mathbf{a} \sim \mathcal{CN}\left(\mathbf{0}_n, \sigma^2 \mathbf{I}_n\right)$ represents a circularly-symmetric complex-valued Gaussian vector $\mathbf{a}$ whose mean and covariance matrix are $\mathbf{0}_n$ and $\sigma^2 \mathbf{I}_n$, respectively. $\mathcal{S} = \{S_{\min}, \ldots, S_{\max}\}$ is an ordered set of integers and $\binom{n}{r}$ denotes the combinations of r objects from a set of n objects. The cardinality of a set $\mathcal{S}$ is denoted as $|\mathcal{S}|$. The empty set is denoted as $\varnothing$. The set of elements in $\mathcal{A}$ but not in $\mathcal{B}$ is denoted as $\mathcal{A} \setminus \mathcal{B}$. Finally, $a^+ := \max(0, a)$ for $a \in \mathbb{R}$.

2.2 System Setup

The SS is comprised by M cells. In each cell, a BS with N antennas and K active single-antenna MSs are located. Thus, there are M BSs, $N_T = M \times N$ transmit antennas, and $K_T = M \times K$ MSs or receive antennas in the cellular network in total. We assume that $K_T \leq M$.

The m-th BS is denoted as BS_m and the k-th MS in the m-th cell is denoted as MS_{km} ($m \in \mathcal{M} = \{1,\ldots,M\}$ and $k \in \mathcal{K} = \{1,\ldots,K\}$). On the other hand, the transmitter and the receiver of the PS are denoted simply as TX_{PS} and RX_{PS}, respectively.

Fig. 2.1 illustrates the system setup [114]. We distinguish between intra-system CCI, which consists of intra-cell MUI and ICI components, and forward / reverse inter-system (FIS / RIS) CCI, as shown in this figure.

The following assumptions are in order:

- All transmissions are narrowband (e.g., corresponding to a subcarrier of a multi-carrier waveform).
- The transport / core network that supports inter-BS cooperation is ideal.
- The nodes have perfect knowledge of the relevant channels. Also, the BSs have knowledge about the interference power threshold (IPT) of the primary receiver(s).
- The transmitted symbols and beamforming (BF) vectors are normalized to unit power.
- Quasi-static frequency-flat standard Rayleigh fading channels and i.i.d. zero-mean additive white Gaussian noise (AWGN) with unit variance are considered.
- The samples of each data signal or noise process are uncorrelated with each other. Also, these random processes are uncorrelated with each other.
- The MSs employ single-user detection, handle the RIS CCI as additional noise, and pass the composite received signal through a whitening filter.

2.3 Signal Models

2.3.1 SISO Primary Channel

Let us consider initially the case where the PS is modeled as a SISO radio link that is established between TX_{PS} and RX_{PS}. The channel between MS_{km} and BS_j is denoted as $\mathbf{h}_{km}^{j} \sim \mathcal{CN}(\mathbf{0}_N, \mathbf{I}_N)$ and its elements correspond to the coefficient of the

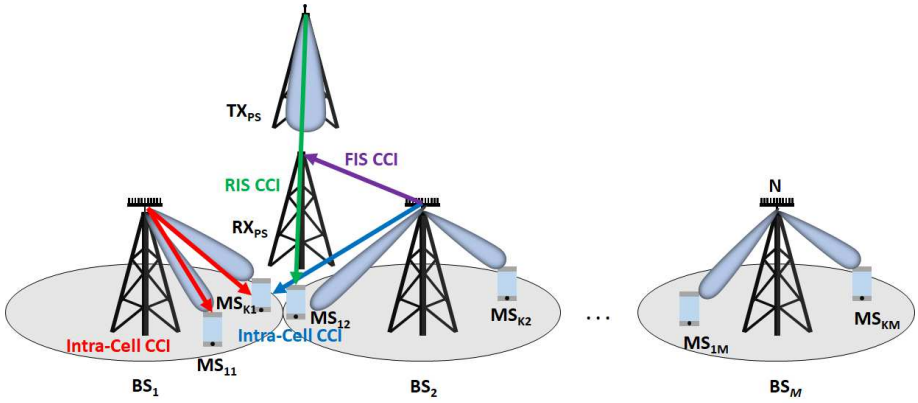


Figure 2.1 System setup, notation, and types of interference for a use case where the PS is a SISO link.

Table 2.1 Channel notation when the PS is a SISO link.

Channel	Receiver	Transmitter
$\mathbf{h}_{km}^{j}$	MS_{km}	BS_j
h_{km}	MS_{km}	TX_{PS}
$\mathbf{g}^{j}$	RX_{PS}	BS_j
g	RX_{PS}	TX_{PS}

channel that couples the receive antenna of MS_{km} with the n-th transmit antenna of BS_j $(n \in \mathcal{N} = \{1,\ldots,N\})$. Similarly, the channel between MS_{km} and TX_{PS} is denoted as $h_{km} \sim \mathcal{CN}(0,1)$. On the other hand, the channel between RX_{PS} and BS_j is denoted as $\mathbf{g}^{j} \sim \mathcal{CN}(\mathbf{0}_N, \mathbf{I}_N)$ and its elements correspond to the coefficient of the channel between the receive antenna of RX_{PS} and the n-th transmit antenna of BS_j. Finally, the channel between RX_{PS} and TX_{PS} is denoted as $g \sim \mathcal{CN}(0,1)$. Channel notation is summarized in Table 2.1.

The BF vector of BS_j that is associated with MS_{km} is denoted as $\mathbf{w}_{mk}^{j} \in \mathbb{C}^N$ and its elements represent the BF weight that is applied at the n-th antenna of BS_j. The power allocated to MS_{km} by BS_j and the symbol transmitted to MS_{km} by BS_j are denoted as $P_{mk}^{j} \in \mathbb{R}_+$ and $s_{mk}^{j} \sim \mathcal{CN}(0,1)$, respectively. Notice that we assume $\left\|\mathbf{w}_{mk}^{j}\right\|^2 = 1$. Similarly, the power allocated to RX_{PS} by TX_{PS} and the symbol transmitted to RX_{PS} by TX_{PS} are denoted as $P \in \mathbb{R}_+$ and $d \sim \mathcal{CN}(0,1)$, respectively. Finally, the AWGN at MS_{km} and RX_{PS} are denoted as $n_{km} \sim \mathcal{CN}(0,1)$ and $z \sim \mathcal{CN}(0,1)$, respectively.

2.3.1.1 Secondary System

At a given transmission slot, the complex baseband representation of the received signal at MS_{km}, $y_{km} \in \mathbb{C}$ ($k \in \mathcal{K}$, $m \in \mathcal{M}$), is [114]:

$$y_{km} = \sum_{j=1}^{M} \sum_{l=1}^{K} \left(\mathbf{h}_{km}^{j}\right)^{\dagger} \mathbf{v}_{jl}^{j} + y_{km}^{(\text{RIS-SISO})} + n_{km}, \tag{2.1}$$

where the precoded signal $\mathbf{v}_{jl}^{j} \in \mathbb{C}^{N}$ is expressed as:

$$\mathbf{v}_{jl}^{j} = \mathbf{w}_{jl}^{j} \sqrt{P_{jl}^{j}} s_{jl}^{j}, \tag{2.2}$$

and $y_{km}^{(\text{RIS-SISO})} \in \mathbb{C}$ is given by:

$$y_{km}^{(\text{RIS-SISO})} = h_{km} \sqrt{P} d. \tag{2.3}$$

Note that the time index has been removed from these equations for better legibility. The first term at the right-hand-side (RHS) of Eq. (2.1) can be expressed as the sum of data, intra-cell MUI, and ICI components as follows [114]:

$$\sum_{j=1}^{M} \sum_{l=1}^{K} \left(\mathbf{h}_{km}^{j}\right)^{\dagger} \mathbf{v}_{jl}^{j} = (\mathbf{h}_{km}^{m})^{\dagger} \mathbf{v}_{mk}^{m} + \sum_{\substack{i=1 \\ i \neq k}}^{K} (\mathbf{h}_{km}^{m})^{\dagger} \mathbf{v}_{mi}^{m} + \sum_{\substack{j=1 \\ j \neq m}}^{M} \sum_{l=1}^{K} \left(\mathbf{h}_{km}^{j}\right)^{\dagger} \mathbf{v}_{jl}^{j}. \tag{2.4}$$

The other two terms at the RHS of Eq. (2.1) represent, from left to right, the RIS CCI and the AWGN at MS_{km}.

2.3.1.2 Primary System

The complex baseband representation of the received signal at RX_{PS}, $y \in \mathbb{C}$, is given by [114]:

$$y = g \sqrt{P} d + \sum_{m=1}^{M} \sum_{k=1}^{K} (\mathbf{g}^{m})^{\dagger} \mathbf{v}_{mk}^{m} + z. \tag{2.5}$$

The terms at the RHS of Eq. (2.5) are, in order, the useful data signal component, the FIS CCI, and the AWGN at RX_{PS}.

2.3.2 MIMO Primary Channel

Now, let us consider the case where TX_{PS} is equipped with $L > 1$ antennas and RX_{PS} is equipped with $Q > 1$ antennas, so that the link between them is characterized as a MIMO channel $\mathbf{G} \in \mathbb{C}^{Q \times L}$ whose entries are i.i.d. complex Gaus-

sian variables $(\mathbf{G})_{ql} \sim \mathcal{CN}(0,1)$ that represent the coefficient of the channel that links the q-th receive antenna of RX_{PS} and the l-th transmit antenna of TX_{PS} ($l \in \mathcal{L} = \{1,\ldots,L\}$). In this scenario, the channel between RX_{PS} and BS_j is modeled by a matrix $\mathbf{G}^j \in \mathbb{C}^{Q\times N}$ whose entries are i.i.d. complex Gaussian variables $\left(\mathbf{G}^j\right)_{qn} \sim \mathcal{CN}(0,1)$ that represent the coefficient of the channel that couples the q-th receive antenna of RX_{PS} with the n-th transmit antenna of BS_j. Also, the channel between MS_{km} and TX_{PS} is denoted as $\mathbf{h}_{km} \in \mathcal{CN}(\mathbf{0}_L, \mathbf{I}_L)$ and its elements represent the coefficient of the channel that links the receive antenna of MS_{km} and the l-th transmit antenna of TX_{PS}.

2.3.2.1 Primary System

The optimal (i.e., capacity-achieving) transmission strategy of the PS in this scenario is precoding / combining based on the SVD of the MIMO channel matrix, $\mathbf{G} = \mathbf{U}\Sigma\mathbf{V}^\dagger$, to decompose the MIMO channel into $r = \text{rank}(\mathbf{G}) = \min(Q,L)$ parallel (i.e., non-interfering) SISO channels (assuming rich scattering conditions) with coefficients its non-zero singular values, $\sigma_i > 0$ ($i = 1,\ldots,r$), followed by the application of the standard WF algorithm to allocate the transmission power P over these channels (or eigenmodes) [8]. Thus, the linear transformations $\mathbf{d} = \mathbf{V}\tilde{\mathbf{d}}$ (pre-processing) and $\tilde{\mathbf{y}} = \mathbf{U}^\dagger\mathbf{y}$ (post-processing) are applied to the vectors of input symbols $\tilde{\mathbf{d}} \sim \mathcal{CN}(\mathbf{0}_L, \mathbf{I}_L)$ and received symbols $\mathbf{y} \in \mathbb{C}^Q$ to produce the vectors of transmitted symbols $\mathbf{d} \sim \mathcal{CN}(\mathbf{0}_L, \mathbf{I}_L)$ and output symbols $\tilde{\mathbf{y}} \in \mathbb{C}^Q$, respectively. The columns of the unitary matrices $\mathbf{U} \in \mathbb{C}^{Q\times Q}$ and $\mathbf{V} \in \mathbb{C}^{L\times L}$ are the left and right singular vectors of $\mathbf{G}$, respectively.

The received signal at RX_{PS} after combining, $\tilde{\mathbf{y}}_s in \mathbb{C}^Q$, is given by [114]:

$$\tilde{\mathbf{y}} = \Sigma\mathbf{P}^{1/2}\tilde{\mathbf{d}} + \sum_{m=1}^{M}\sum_{k=1}^{K}\mathbf{U}^\dagger\mathbf{G}^m\mathbf{v}_{mk}^m + \tilde{\mathbf{z}}, \tag{2.6}$$

where $\tilde{\mathbf{z}} = \mathbf{U}^\dagger\mathbf{z}$, with $\mathbf{z} \sim \mathcal{CN}(\mathbf{0}_Q, \mathbf{I}_Q)$ and $\tilde{\mathbf{z}} \sim \mathcal{CN}(\mathbf{0}_Q, \mathbf{I}_Q)$ representing the AWGN before and after receive combining, respectively. $\Sigma \in \mathbb{C}^{Q\times L}$ holds the singular values of $\mathbf{G}$ in decreasing order, i.e., $(\Sigma)_{ii} = \sigma_i$ with $\sigma_1 \geq \cdots \geq \sigma_i \geq \cdots \geq \sigma_r$ and $(\Sigma)_{ql} = 0$ for $q \neq l$. Finally, $\mathbf{P} = \text{diag}(P_1,\ldots,P_r,0,\ldots,0) \in \mathbb{C}^{L\times L}$ is the PA matrix. As in Eq. (2.5), the first term at the RHS of Eq. (2.6) is the useful data signal component that is received by RX_{PS}, whereas the second term is the received FIS CCI component.

2.3.2.2 Secondary System

The RIS CCI component of the received signal at MS_{km} ($k \in \mathcal{K}$, $m \in \mathcal{M}$) is given by [114]:

$$y_{km}^{(\text{RIS-MIMO})} = \sqrt{P}\left(\mathbf{h}_{km}\right)^\dagger\mathbf{d} = \sqrt{P}\left(\mathbf{h}_{km}\right)^\dagger\mathbf{V}\tilde{\mathbf{d}}. \tag{2.7}$$

2.3.3 MIMO Broadcast Primary Channel

Next, we consider the case where TX_{PS} serves $Q \geq 2$ single-antenna primary receivers RX_q over a MIMO BC and is equipped with $L \geq Q$ antennas. The channel between RX_q and TX_{PS} is denoted as $\mathbf{h}_q \sim \mathcal{CN}(\mathbf{0}_L, \mathbf{I}_L)$ and its elements represent the coefficient of the channel that couples the receive antenna of RX_q with the l-th transmit antenna of TX_{PS}. Similarly, the channel between RX_q and BS_j is denoted as $\mathbf{g}_q^j \sim \mathcal{CN}(\mathbf{0}_N, \mathbf{I}_N)$ and its elements represent the coefficient of the channel that links the receive antenna of RX_q and the n-th transmit antenna of BS_j. The BF vector associated with RX_q by TX_{PS} is denoted as $\mathbf{w}_q \in \mathbb{C}^L$ and its elements represent the corresponding BF weight that is applied at the l-th antenna of TX_{PS}. Notice that $\left\|\mathbf{w}_q\right\|^2 = 1$. The power allocated to RX_q by TX_{PS} and the symbol transmitted to RX_q by TX_{PS} are denoted as $P_q \in \mathbb{R}_+$ and $d_q \sim \mathcal{CN}(0,1)$, respectively. Finally, $z_q \sim \mathcal{CN}(0,1)$ denotes the AWGN at RX_q.

2.3.3.1 Primary System

Assuming the application of linear precoding, the received signal at RX_q, $y_q \in \mathbb{C}$, is given by [114]:

$$y_q = \left(\mathbf{h}_q\right)^\dagger \mathbf{v}_q + \sum_{\substack{i=1 \\ i \neq q}}^{Q} \left(\mathbf{h}_q\right)^\dagger \mathbf{v}_i + \sum_{m=1}^{M} \sum_{k=1}^{K} \left(\mathbf{g}_q^m\right)^\dagger \mathbf{v}_{mk}^m + z_q, \quad q \in \mathcal{Q}, \tag{2.8}$$

where $\mathbf{v}_q \in \mathbb{C}^L$ is expressed as [114]:

$$\mathbf{v}_q = \mathbf{w}_q \sqrt{P_q} d_q. \tag{2.9}$$

The terms at the RHS of Eq. (2.8) are, from left to right, the useful data component, the inter-user interference, the FIS CCI, and the AWGN at RX_q.

Let us ignore, for simplicity, the FIS CCI in Eq. (2.8). Then, we can rewrite this equation as [114]:

$$y_q = \sum_{i=1}^{Q} \left(\mathbf{h}_q\right)^\dagger \mathbf{w}_i \sqrt{P_i} d_i + z_q. \tag{2.10}$$

By stacking together all the received symbols, transmitted symbols, and noise samples into vectors $\mathbf{y}_{\text{PS}} \in \mathbb{C}^Q$, $\mathbf{d}_{\text{PS}} \in \mathbb{C}^Q$, and $\mathbf{z}_{\text{PS}} \in \mathbb{C}^Q$, respectively, we obtain the composite system model [114]:

$$\mathbf{y}_{\text{PS}} = \mathbf{H}_{\text{PS}} \mathbf{W}_{\text{PS}} \mathbf{P}_{\text{PS}}^{1/2} \mathbf{d}_{\text{PS}} + \mathbf{z}_{\text{PS}}, \tag{2.11}$$

where $\mathbf{H}_{\text{PS}} \in \mathbb{C}^{Q \times L}$ is the composite channel matrix whose q-th row holds the channel of TX_{PS} with RX_q, $\left(\mathbf{h}_q\right)^\dagger \in \mathbb{C}^L$; $\mathbf{W}_{\text{PS}} \in \mathbb{C}^{L \times Q}$ is the precoding matrix whose q-th column holds the BF vector for RX_q, $\mathbf{w}_q \in \mathbb{C}^L$; and $\mathbf{P}_{\text{PS}} = \text{diag}\left(P_1, \ldots, P_Q\right) \in \mathbb{R}^{Q \times Q}$ is the PA matrix.

2.3.3.2 Secondary System

The RIS CCI component of the received signal at MS_{km} $(k \in \mathcal{K},\ m \in \mathcal{M})$ is given by [114]:

$$y_{km}^{(\text{RIS-MIMO-BC})} = \left(\mathbf{h}_{km}\right)^\dagger \sum_{q=1}^{Q} \mathbf{v}_q. \tag{2.12}$$

2.4 Coordinated Power Allocation

2.4.1 Instantaneous SINR

The instantaneous signal-to-interference-plus-noise-ratio (SINR) at MS_{km} is given by [114]:

$$\gamma_{km} = \frac{\left|\left(\mathbf{h}_{km}^m\right)^\dagger \mathbf{v}_{mk}^m\right|^2}{\sum\limits_{\substack{i=1 \\ i \neq k}}^{K} \left|\left(\mathbf{h}_{km}^m\right)^\dagger \mathbf{v}_{mi}^m\right|^2 + \sum\limits_{\substack{j=1 \\ j \neq m}}^{M} \sum_{l=1}^{K} \left|\left(\mathbf{h}_{km}^j\right)^\dagger \mathbf{v}_{jl}^j\right|^2 + I_{km}}, \tag{2.13}$$

where

$$\left|\left(\mathbf{h}_{km}^m\right)^\dagger \mathbf{v}_{mk}^m\right|^2 = \left|\left(\mathbf{h}_{km}^m\right)^\dagger \mathbf{w}_{mk}^m\right|^2 P_{mk}^m \tag{2.14}$$

and I_{km} is the sum of the noise power and the power of the RIS CCI component. The other terms in the denominator of Eq. (2.13) are, from left to right, the power of intra-cell MUI and of ICI, while the nominator corresponds to the power of the data signal component.

2.4.1.1 SISO Primary Channel

When the PS is a SISO link, then:

$$I_{km} = \left|h_{km}\right|^2 P + 1. \tag{2.15}$$

2.4.1.2 MIMO Primary Channel

When the PS is comprised by a MIMO link, we obtain:

$$I_{km} = \left\|\left(\mathbf{h}_{km}\right)^\dagger \mathbf{V}\right\|^2 P + 1. \tag{2.16}$$

2.4.1.3 MIMO Broadcast Primary Channel

When the PS refers to a MIMO BC, we have:

$$I_{km} = \left\| (\mathbf{h}_{km})^{\dagger} \mathbf{W}_{\mathrm{PS}} \right\|^2 P + 1. \tag{2.17}$$

2.4.2 Instantaneous Rate

Given that the transmitted symbols are i.i.d. zero-mean complex Gaussian random variables, the instantaneous bandwidth-normalized data rate or spectral efficiency (SE) of MS_{km}, $R_{km} \in \mathbb{R}_+$, is given by the Shannon formula:

$$R_{km} = \log_2 (1 + \gamma_{km}). \tag{2.18}$$

2.4.3 Instantaneous Sum-Rate

The instantaneous bandwidth-normalized sum-rate (SR) capacity, i.e., the instantaneous sum-SE, is given by:

$$R = \sum_{m=1}^{M} \sum_{k=1}^{K} R_{km}. \tag{2.19}$$

2.4.4 Transmission Constraints

2.4.4.1 Transmission Power Constraints

Each transmission from a BS to one of its K users has non-negative power:

$$P_{mk}^{m} \geq 0. \tag{2.20}$$

Also, the transmissions of each BS to its K users are subject to a sum-power constraint (SPC). That is, the total transmission power should not exceed a maximum value P_T [114]:

$$\sum_{k=1}^{K} P_{mk}^{m} \leq P_T, \quad m \in M. \tag{2.21}$$

2.4.4.2 Interference Power Constraints

The operation of the SS is subject to an interference power constraint (IPC) per primary receiver, which states that the total power of the FIS CCI that is received by this primary receiver should not exceed an IPT P_I.

For the cases where the PS is a SISO link or a MIMO one, we have [114]:

$$\sum_{m=1}^{M}\sum_{k=1}^{K}\alpha_{mk}^{m}P_{mk}^{m}\leq P_{I}, \tag{2.22}$$

whereas in the scenario where the PS is a MIMO BC with Q single-antenna PUs, there are Q IPCs of the form [114]:

$$\sum_{m=1}^{M}\sum_{k=1}^{K}\left(\alpha_{mk}^{m}\right)^{(q)}P_{mk}^{m}\leq P_{I},\quad q\in\mathcal{Q}. \tag{2.23}$$

SISO Primary Channel: When the PS is a SISO link, we have:

$$\alpha_{mk}^{m}=\left|\left(\mathbf{g}^{m}\right)^{\dagger}\mathbf{w}_{mk}^{m}\right|^{2}. \tag{2.24}$$

MIMO Primary Channel: When the PS is a MIMO link, we define:

$$\alpha_{mk}^{m}=\|\mathbf{G}^{m}\mathbf{w}_{mk}^{m}\|^{2}. \tag{2.25}$$

MIMO Broadcast Primary Channel: Finally, when the PS is a MIMO BC:

$$\left(\alpha_{mk}^{m}\right)^{(q)}=\left|\left(\mathbf{g}_{q}^{m}\right)^{\dagger}\mathbf{w}_{mk}^{m}\right|^{2}. \tag{2.26}$$

2.4.4.3 QoS Constraints

In some cases, we should ensure that the instantaneous data rate of MS_{km} is at least equal to a minimum value $\tilde{R}_{km}>0$ (e.g., in streaming video applications [8]). These K_T MRCs are expressed as:

$$R_{km}\geq\tilde{R}_{km}. \tag{2.27}$$

In view of Eq. (2.18), we can represent these constraints as [114]:

$$\gamma_{km}\geq\tilde{\gamma}_{km}, \tag{2.28}$$

where $\tilde{\gamma}_{km}$ is the minimum required SINR of MS_{km}.

2.4.5 Coordinated ZF Precoding

Let us consider the application of coordinated ZF precoding at the SS. The precoding matrix for the m-th BS, $\mathbf{W}_m \in \mathbb{C}^{M \times K_T}$, is given by [114]:

$$\mathbf{F}_m^{(\mathrm{ZF})} = \mathbf{H}_m^{\#} = \mathbf{H}_m^{\dagger} \left(\mathbf{H}_m \mathbf{H}_m^{\dagger} \right)^{-1}. \tag{2.29a}$$

$$\left(\mathbf{W}_m^{(\mathrm{ZF})} \right)_{*j} = \frac{\left(\mathbf{F}_m^{(\mathrm{ZF})} \right)_{*j}}{\left\| \left(\mathbf{F}_m^{(\mathrm{ZF})} \right)_{*j} \right\|}, \quad j = 1, \ldots, K_T. \tag{2.29b}$$

In Eq. (2.29a), $\mathbf{H}_m \in \mathbb{C}^{K_T \times M}$ is a concatenated matrix defined as

$$\mathbf{H}_m = \begin{bmatrix} X_1^T & \ldots & X_{M'}^T, \end{bmatrix}^T \tag{2.30}$$

where $\mathbf{X}_1 = \left[\left(\mathbf{h}_{11}^m \right)^T \quad \cdots \quad \left(\mathbf{h}_{K1}^m \right)^T \right]^T$. That is, $\mathbf{H}_m$ holds the channels between BS_m and all users in all cells. ZF precoding completely eliminates the intra-SS CCI. Thus, the SINR of MS_{km} after the application of ZF precoding becomes:

$$\left(\gamma_{km} \right)^{(\mathrm{ZF})} = \frac{\left| \left(\mathbf{h}_{km}^m \right)^{\dagger} \left(\mathbf{w}_{mk}^m \right)^{(\mathrm{ZF})} \right|^2 P_{mk}^m}{I_{km}}. \tag{2.31}$$

2.4.6 Coordinated Power Allocation Problems

We consider the determination of the PA scheme that maximizes the SR of the SS under the aforementioned transmission constraints, assuming the application of C-ZF precoding. This optimization problem, **P1**, takes the form [114]:

$$\min_{\substack{P_{mk}^m \\ m \in \mathcal{M},\, k \in \mathcal{K}}} \quad -R = -\sum_{m=1}^{M} \sum_{k=1}^{K} \log_2 \left(1 + \lambda_{mk}^m P_{mk}^m \right) \tag{2.32a}$$

s.t.

$$\sum_{k=1}^{K} P_{mk}^m \leq P_T, \quad m \in \mathcal{M}, \tag{2.32b}$$

$$\sum_{m=1}^{M} \sum_{k=1}^{K} \alpha_{mk}^m P_{mk}^m \leq P_I, \tag{2.32c}$$

$$P_{mk}^m \geq \tilde{P}_{mk}^m, \quad k \in \mathcal{K},\ m \in \mathcal{M}, \tag{2.32d}$$

where $\lambda_{mk}^m = \gamma_{km} / P_{mk}^m$ and the per-user QoS constraints in Eq. (2.32d) are derived from Eq. (2.28) by substituting $\gamma_{km} = \lambda_{mk}^m P_{mk}^m$ and $\tilde{\gamma}_{km} = \lambda_{mk}^m \tilde{P}_{mk}^m$, with $\tilde{P}_{mk}^m$ denoting

Table 2.2 Coordinated power allocation problems.

Problem	Description
P1	SR Maximization under SPCs, MRCs, and IPC(s).
P2	SR Maximization under SPCs and IPC(s).
P3	SR Maximization under SPCs.
P4	SR Maximization under SPCs and MRCs.

the minimum power that should be allocated to MS_{km}. By setting $\tilde{P}^m_{mk} = 0$, the per-user QoS constraints in Eq. (2.32d) are converted into the non-negative allocated power per-user constraints of Eq. (2.20) and the resulting optimization problem is referred to as **P2**. If, in addition, we omit the IPC of Eq. (2.32c), we obtain **P3**. Finally, if we omit the IPC(s) in **P1** we obtain **P4**. The description of these optimization problems is summarized in Table 2.2. Notice that these PA tasks are convex, since the use of ZF precoding has removed the coupled interference terms from the SINR.

2.4.7 Optimal Coordinated Power Allocation Schemes

The solutions to **P1**–**P4** are presented in Theorem 2.1 [114, 115].

Theorem 2.1 *Assuming that the PS is a SISO or MIMO link, the solution to **P1** is given by the coordinated QoS-aware interference-constrained PA (CQA-ICPA) scheme:*

$$P^m_{mk} = \left(\frac{1}{\ln 2 \left(\nu_m + \mu \alpha^m_{mk} \right)} - \frac{1}{\lambda^m_{mk}} - \tilde{P}^m_{mk} \right)^+ + \tilde{P}^m_{mk}, \tag{2.33}$$

where ν_m and μ are Lagrange multipliers associated with the transmit power constraints and the IPC, respectively.

*Similarly, the solution to **P2** is the coordinated ICPA (C-ICPA) scheme that is obtained from Eq. (2.33) by setting $\tilde{P}^m_{mk} = 0$ (i.e., by deactivating the MRCs):*

$$P^m_{mk} = \left(\frac{1}{\ln 2 \left(\nu_m + \mu \alpha^m_{mk} \right)} - \frac{1}{\lambda_{km}} \right)^+. \tag{2.34}$$

*In addition, the solution to **P3** is the standard coordinated water-filling PA or coordinated interference-unconstrained PA (C-IUPA) used in isolated CoMP setups that is obtained from Eq. (2.34) by setting $\mu = 0$ (i.e., by deactivating the IPC):*

$$P^m_{mk} = \left(\frac{1}{\ln 2 \nu_m} - \frac{1}{\lambda_{km}} \right)^+. \tag{2.35}$$

*Finally, the solution to **P4** is the coordinated QoS-aware IUPA (CQA-IUPA), which is obtained by Eq. (2.33) by setting $\mu = 0$ (i.e., it corresponds to a standalone CoMP setup that operates under MRCs):*

$$P_{mk}^{m} = \left(\frac{1}{\ln 2 \nu_m} - \frac{1}{\lambda_{mk}^{m}} - \tilde{P}_{mk}^{m} \right)^{+} + \tilde{P}_{mk}^{m}, \tag{2.36}$$

When the PS is a MIMO BC, $\alpha_{mk}^{m}\mu$ in Eq. (2.33) and Eq. (2.34) is replaced by $\sum_{q=1}^{Q} \left(\alpha_{mk}^{m}\right)^{(q)} \mu_q$, where μ_q are the Lagrange multipliers for the Q IPCs ($q \in \mathcal{Q}$).

Proof: These solutions are obtained by taking the Lagrangian form of the corresponding optimization problems and applying the Karush-Kuhn-Tucker (KKT) conditions [116]. The proofs are given in Appendix A.

2.4.8 Power Allocation Algorithm

The iterative algorithm that calculates the Lagrange multipliers $\nu^* = \left[\nu_1^* \cdots \nu_M^*\right]$ and μ^* and implements the coordinated PA schemes described in Section 2.4.7, under the assumption that the PS is either a SISO or a MIMO link, is presented in Algorithm 2.1. The algorithm, whose accuracy is controlled by the parameter $\delta_\mu > 0$, makes use of the bisection method to update the value of μ in each iteration based on whether the IPC is met or not [114]. Note that when the IPC is inactive, $\mu = 0$ and the algorithm reduces to the corresponding IUPA solution [5].

For the scenario where the PS is a MIMO BC, Algorithm 2.1 should be modified to search for the optimal ν^* and $\mu^* = \left[\mu_1^* \cdots \mu_Q^*\right]$. The ellipsoid method, which is a generalization of the one-dimensional bisection method for higher dimensions, can be used to update simultaneously all μ_q^* ($q \in \mathcal{Q}$) [117]. Alternatively, a subgradient based method, which converges rapidly even for a large number of users and has low computational complexity, may be used [117].

2.4.9 Heuristic Power Allocation

Linear precoding schemes try to balance between the mitigation of the inter-user interference and the increase of the receive signal-to-noise-ratio (SNR) at the intended users [5, 51]. The determination of the optimum linear precoding scheme is for many problems of interest, such as the (W)SR maximization, computationally prohibitive [5, 51]. Thus, we commonly rely on simple heuristics.

ZF precoding is the most representative example. In this transmission strategy, the BF vectors are orthogonal to other active users' channel vectors to eliminate the CCI [5]. ZF precoding is asymptotically optimal with the number of users in the interference-limited high SNR regime when single-antenna MSs are utilized but it performs poorly in the noise-limited low SNR regime [5].

Algorithm 2.1 CQA-ICPA and C-ICPA algorithm for CP.

procedure CQA-ICPA($\lambda_{mk}^{m}, \alpha_{mk}^{m}, P_T, P_I, \tilde{P}_{mk}^{m}$)
  Initialize: $\mu_{\min}$, $\mu_{\max}$
  while $|\mu_{\max} - \mu_{\min}| > \delta_\mu$ **do**
    if (C-ICPA) **then**
      Set: $\tilde{P}_{mk}^{m} = 0$
    end if
    $\mu = (\mu_{\min} + \mu_{\max})/2$
    for $m = 1$ to M **do**
      Find $\min(\nu_m)$, $\nu_m \geq 0$:
    $\sum_{k=1}^{K} \left(\left(\frac{1}{\ln 2 \left(\nu_m + \mu \alpha_{mk}^{m} \right)} - \frac{1}{\lambda_{mk}^{m}} - \tilde{P}_{mk}^{m} \right)^{+} + \tilde{P}_{mk}^{m} \right) \leq P_T$
    end for
    Compute P_{mk}^{m} according to Eq. (2.33) (CQA-ICPA) or Eq. (2.34) (C-ICPA)
    if $\sum_{m=1}^{M} \sum_{k=1}^{K} \alpha_{mk}^{m} P_{mk}^{m} \geq P_I$ **then**
      $\mu_{\min} = \mu$
    else
      $\mu_{\max} = \mu$
    end if
  end while
  Output: $P_{mk}^{m}, \quad m \in \mathcal{M};\ k \in \mathcal{K}$
end procedure

In maximum ratio transmission (MRT), on the other hand, the BF vector of each user matches to its channel vector to maximize the receive SNR [5]:

$$\left(\tilde{\mathbf{w}}_{mk}^{m}\right)^{(\mathrm{MRT})} = \mathbf{h}_{km}^{m}. \tag{2.37a}$$

$$\left(\mathbf{w}_{mk}^{m}\right)^{(\mathrm{MRT})} = \frac{\left(\tilde{\mathbf{w}}_{mk}^{m}\right)^{(\mathrm{MRT})}}{\left\| \left(\tilde{\mathbf{w}}_{mk}^{m}\right)^{(\mathrm{MRT})} \right\|}. \tag{2.37b}$$

In contrast to ZF precoding, MRT performs well in the noise-limited low-SNR regime, since it focuses the radiated power towards the intended users, and is the optimal strategy in this SNR regime when a single user is scheduled at each slot. On the other hand, its capacity floors in the interference-limited high-SNR regime due to the uncoordinated CCI [5].

Regularized ZF (RZF) is an extension of ZF precoding that improves the performance in the low-SNR regime [5]:

$$\mathbf{F}_{m}^{(\mathrm{RZF})} = \mathbf{H}_{m}^{\dagger} \left(\frac{1}{\alpha_{req}} \mathbf{I}_{K_T} + \mathbf{H}_m \mathbf{H}_{m}^{\dagger} \right)^{-1}. \tag{2.38a}$$

$$\mathbf{W}_{m}^{(\mathrm{RZF})} = \frac{\left(\mathbf{F}_{m}^{(\mathrm{RZF})}\right)_{*j}}{\left\| \left(\mathbf{F}_{m}^{(\mathrm{RZF})}\right)_{*j} \right\|}, \tag{2.38b}$$

where $j = 1, \ldots, K_T$ and $\alpha_{req} = 1/MP_T$ is the regularization factor [118]. Other values of α_{req} are also possible [5].

Although these precoding techniques do not eliminate the intra-system CCI within the SS, we can apply heuristically the PA solutions derived in this section for ZF precoding [5].

2.4.10 Coordinated Interference-Constrained Equal PA

A simple suboptimal PA method is coordinated interference-constrained equal power allocation (C-ICEPA), which allocates equal powers to the users, taking though into account both the SPCs and the IPC, as shown in Proposition 2.1 [114].

Proposition 2.1 *The coordinated interference-constrained equal PA (C-ICEPA) scheme allocates the following power to each user:*

$$P_{mk}^{m} = \begin{cases} \min\left(\frac{P_T}{K}, \frac{P_I}{\sum_{m=1}^{M}\sum_{k=1}^{K}\alpha_{mk}^{m}}\right) & \text{SISO / MIMO} \\ \min\left(\frac{P_T}{K}, \frac{P_I}{\sum_{q=1}^{Q}\sum_{m=1}^{M}\sum_{k=1}^{K}\left(\alpha_{mk}^{m}\right)^{(q)}}\right) & \text{MIMO BC} \end{cases}, \tag{2.39}$$

where the terms SISO, MIMO, and MIMO BC refer to the PS setup and P_T/K are the power levels that are allocated to the users in coordinated interference-unconstrained equal power allocation (C-IUEPA).

Remark *Note that for $P_I \to \infty$, $P_{mk}^{m} \to P_T/K$, whereas for $P_I \to 0$, $P_{mk}^{m} \to 0$.*

2.5 Coordinated Projected Zero-Forcing Precoding

When the IPT is extremely hard, we can simply assume as well that the PS does not tolerate any FIS CCI.

2.5.1 SISO Primary Channel

Let us consider initially the case where the PS is a SISO link. The channel between $\mathrm{RX_{PS}}$ and BS_m is $\mathbf{g}^m \in \mathbb{C}^N$ $(m \in \mathcal{M})$. We define $\hat{\mathbf{g}}_m \in \mathbb{C}^N$ as [114]:

$$\hat{\mathbf{g}}_m = \frac{(\mathbf{g}_m)^{\dagger}}{\|\mathbf{g}_m\|}. \tag{2.40}$$

In this scenario, instead of computing the ZF precoder based on the composite channel matrix $\mathbf{H}_m$, we calculate it based on the projection of this matrix into the null space of $(\mathbf{g}_m)^\dagger$, $\hat{\mathbf{H}}_m \in \mathbb{C}^{K_T \times N}$ [114]:

$$\hat{\mathbf{H}}_m = \mathbf{H}_m \left(\mathbf{I}_N - \hat{\mathbf{g}}_m \left(\hat{\mathbf{g}}_m \right)^\dagger \right). \tag{2.41}$$

That is, we define the precoder $\mathbf{W}_m^{(\text{P-ZF})}$ as [114]:

$$\mathbf{F}_m^{(\text{P-ZF})} = \hat{\mathbf{H}}_m^{\#} = \hat{\mathbf{H}}_m^\dagger \left(\hat{\mathbf{H}}_m \hat{\mathbf{H}}_m^\dagger \right)^{-1}. \tag{2.42a}$$

$$\mathbf{W}_m^{(\text{P-ZF})} = \frac{\left(\mathbf{F}_m^{(\text{P-ZF})} \right)_{*j}}{\left\| \left(\mathbf{F}_m^{(\text{P-ZF})} \right)_{*j} \right\|}, \quad j = 1, \ldots, K_T. \tag{2.42b}$$

This precoding scheme is called coordinated projected ZF (P-ZF) and completely eliminates the FIS CCI at the primary receiver. Indeed, if we ignore, for convenience, the normalization of the precoding matrix, we have $\hat{\mathbf{H}}_m \mathbf{F}_m^{(\text{P-ZF})} = \hat{\mathbf{H}}_m \hat{\mathbf{H}}_m^{\#} = \mathbf{I}_N$.

Remark *When the IPT is null, i.e., $(P_I = 0)$, coordinated P-ZF is the optimum transmission strategy.*

On the other hand, P-ZF precoding does not cancel the intra-SS CCI. However, similar to the approach used in Section 2.4.9 and since the FIS CCI has been removed, we can use heuristically the C-IUPA scheme (or the CQA-IUPA method, if we should also meet some given MRCs).

2.5.2 MIMO Primary Channel

Next, let us consider the case where the PS is a MIMO link. The channel between RX_{PS} and BS_m is $\mathbf{G}^m \in \mathbb{C}^{Q \times N}$ $(m \in \mathcal{M})$. We define $\hat{\mathbf{G}}_m \in \mathbb{C}^{N \times Q}$ as [114]:

$$\hat{\mathbf{G}}_m = \frac{(\mathbf{G}_m)^\dagger}{\|\mathbf{G}_m\|}. \tag{2.43}$$

Then, $\hat{\mathbf{H}}_m \in \mathbb{C}^{K_T \times N}$ is defined as [114]:

$$\hat{\mathbf{H}}_m = \mathbf{H}_m \left(\mathbf{I}_N - \hat{\mathbf{G}}_m \left(\hat{\mathbf{G}}_m \right)^\dagger \right), \tag{2.44}$$

where $\hat{\mathbf{G}}_m \left(\hat{\mathbf{G}}_m \right)^\dagger \in \mathbb{C}^{N \times N}$. The (non-normalized) precoding matrix is $\mathbf{F}_m^{(\text{P-ZF})} = \hat{\mathbf{H}}_m^{\#}$.

2.5.3 MIMO Broadcast Primary Channel

Let us denote as $\mathbf{G}^m \in \mathbb{C}^{Q \times N}$ the composite channel between BS_m and the Q single-antenna primary receivers. Furthermore, let $\mathbf{G}_m = \mathbf{Q}_m \Lambda_m^{1/2} \mathbf{U}_m^\dagger$ be the SVD of $\mathbf{G}_m$. Then, we obtain $\hat{\mathbf{H}}_m$ as the projection of $\mathbf{H}_m$ onto the null space of $\mathbf{G}_m^\dagger$ [114]:

$$\hat{\mathbf{H}}_m = \mathbf{H} \left(\mathbf{I}_N - \mathbf{U}_m \mathbf{U}_m^\dagger \right). \tag{2.45}$$

Note that this matrix projection is non-trivial only when $N > Q$; otherwise, $\hat{\mathbf{H}}_m$ is a null matrix. The (non-normalized) precoding matrix is $\mathbf{F}_m^{(\text{P-ZF})} = \hat{\mathbf{H}}_m^{\#}$.

2.6 Heuristic Coordinated User Selection

Let us assume that $U > N$ single-antenna MSs (users) are requesting service during each scheduling slot in each cell. There are $O = \binom{U}{K}$ possible subsets of K users in the m-th cell $\mathcal{U}_m^{(o)}$ ($o \in \mathcal{O} = \{1, \ldots, O\}$, $m \in \mathcal{M} = \{1, \ldots, M\}$), thus resulting in $S = O^M$ possible subsets of $K_T = MK$ users in the network / cluster $\mathcal{U}^{(s)}$ ($s \in \mathcal{S} = \{1, \ldots, S\}$). A central scheduler selects one of these subsets $\mathcal{U}^{(s)}$, $\left(\mathcal{U}^{(s)}\right)^*$, according to the given performance metric and transmission constraints.

2.6.1 Problem Statement

The maximum SR for a set of users $\mathcal{U}^{(s)}$ that is obtained via the coordinated precoding and power allocation techniques described in the previous sections is denoted as $\left(R^{(s)}\right)^*$ ($s \in \mathcal{S}$). The maximum of the S obtained values $\left(R^{(s)}\right)^*$ is denoted as R^* [114]:

$$R^* = \max_{s \in \mathcal{S}} \left(R^{(s)}\right)^* \tag{2.46}$$

and the corresponding user set is denoted as $\mathcal{U}^*$.

Our goal is to determine the set of active users $\mathcal{U}^* \in \Omega$, where Ω represents the collection of all sets $\mathcal{U}^{(s)}$, as the candidate user set that corresponds to the maximum achievable SR [114]:

$$\mathcal{U}^* = \arg \max_{\mathcal{U}^{(s)} \in \Omega} \left(R^{(s)}\right)^*. \tag{2.47}$$

Note that the optimal solution for this combinatorial problem is obtained via exhaustive search, whose computational complexity is prohibitively high for multi-cell multi-user setups. Hence, we have to rely on suboptimal alternatives.

2.6.2 Reduced Search Space User Selection

In reduced search space (RSS) user selection, the user scheduler *randomly selects prior to the precoding and power allocation processes* $S' < S$ *candidate user sets of* $K_T = MK$ *users,* $\left(\mathcal{U}^{(s')}\right)'$ $(s' \in \mathcal{S}' = \{1,\ldots,S'\})$, out of the S candidate user sets $\mathcal{U}^{(s)}$ $(s \in \mathcal{S})$. That is, it randomly selects $O' < O$ candidate user sets of K users in each cell, $\left(\mathcal{U}_m^{(o')}\right)'$ $(o' \in \mathcal{O}' = \{1,\ldots,O'\},\ m \in \mathcal{M})$, out of the O candidate user sets $\mathcal{U}_m^{(o)}$ $(o \in \mathcal{O})$. Then, the precoders and power levels are computed only for these S' candidate user sets. Finally, the user scheduler performs exhaustive search over this limited number of candidate user sets to determine the set of active users, $(\mathcal{U}^*)'$, i.e., the set of users that maximizes the SR [114]:

$$(\mathcal{U}^*)' = \arg \max_{\left(\mathcal{U}^{(s')}\right)' \in \Omega'} \left(R^{(s')}\right)^*, \tag{2.48}$$

where Ω' is the collection of all the considered sets $\left(\mathcal{U}^{(s')}\right)'$. The design parameter S' determines the computational complexity of this user selection method. The performance benchmark is the optimal user selection, where the precoders and power levels are computed for all S candidate user sets and user scheduling is based on exhaustive search over this search space.

2.6.2.1 Greedy-Like Implementation

Note that RSS user scheduling can be performed in a "greedy-like" manner. According to this approach, we first select randomly one of the candidate user sets S' and then we select (again randomly) one user from this set in each cell. This set of users $\mathcal{A}_0$ is our initial selected set $\mathcal{F}$. From this starting point, we can end up in a number of possible user sets, which is, however, a subset of S'. This is because the initially selected M-tuple of users is not a member of all S' candidate user sets, but only of some of them. The collection of all valid sets where we can end up after the initialization is denoted as V. For the initially selected M-tuple of users, we compute the achieved sum-SE R_0, assuming the use of the considered CP and coordinated PA schemes under consideration. Next, we add randomly a user in each cell, taking though into account which users we can add (i.e., considering the subset of valid user candidate sets, according to the aforementioned initialization), and we compute again the achieved sum-SE R_1 for these $2M$ users (recall that we have M cells with 2 users per cell in this step) of this new set $\mathcal{A}_1$ formed after the first iteration of the user scheduling algorithm. If the sum-SE increases (i.e., if $R_1 > R_0$), then we set $\mathcal{A}_1$ as our currently selected set $\mathcal{F}$, we add a valid user in each cell to form the set $\mathcal{A}_2$, and repeat the procedure. Otherwise, we randomly discard one of the users added in the last step, replace her / him with another

Algorithm 2.2 Greedy RSS User Selection.

procedure GRSS(M, K)
  Initialize: $\mathcal{A}_0$, V
  Set: $\mathcal{F} = \mathcal{A}_0$
  while $|\mathcal{F}| < K$ **do**
    Form: $\mathcal{A}_k \in V$
    Compute: R_k
    if $R_k > R_{k-1}$ **then**
      Set: $\mathcal{F} = \mathcal{A}_k$
      Form: $\mathcal{A}_{k+1} \in V$
    else
      Replace a user in $\mathcal{A}_k \in V$ and repeat
    end if
  end while
  Output: $\mathcal{F}$
end procedure

valid user, recompute the sum-SE R_1, compare it with R_0, and act accordingly (i.e., either set this user set as the new selected set if $R_1 > R_0$ or discard again one of the added users and replace it with another valid user if $R_1 < R_0$). This procedure continuous until we have K users in each cell. Clearly, the minimum number of iterations is K. This greedy RSS (GRSS) user selection algorithm is summarized in Algorithm 2.2.

2.6.3 Inter-System Correlation-Aware User Selection

Let us define the cross-correlation between each direct user channel $\left(\mathbf{h}_{km}^{m}\right)^{(s)}$ of the user set $\mathcal{U}_m^{(s)}$ $(k \in \mathcal{K},\ m \in \mathcal{M})$ and each one of the FIS CCI channels.

2.6.3.1 SISO Primary Channel

When the PS is a SISO link, we have [114]:

$$\left(\rho_{km}^{j}\right)^{(s)} = \frac{\left(\left(\mathbf{h}_{km}^{m}\right)^{(s)}\right)^{\dagger}\mathbf{g}^{j}}{\left\|\left(\mathbf{h}_{km}^{m}\right)^{(s)}\right\|\left\|\mathbf{g}^{j}\right\|}, \tag{2.49}$$

where $\mathbf{g}^j$ is the channel between $\mathrm{RX_{PS}}$ and BS_j $(j \in \mathcal{M})$. Since there are $K_T = MK$ direct user channels and M FIS CCI channels, we obtain $n_T = MK_T = M^2K$ such values.

2.6.3.2 MIMO Primary Channel

When the PS is a MIMO link, we compare each direct user channel $\left(\mathbf{h}_{km}^{m}\right)^{(s)}$ with the FIS CCI channel between BS_j and the q-antenna of $\mathrm{RX}_{\mathrm{PS}}$ $(q \in \mathcal{Q})$ [114]:

$$\left(\rho_{km}^{j,q}\right)^{(s)} = \frac{\left(\left(\mathbf{h}_{km}^{m}\right)^{(s)}\right)^{\dagger}\left(\mathbf{G}^{j}\right)_{q*}}{\left\|\left(\mathbf{h}_{km}^{m}\right)^{(s)}\right\|\left\|\left(\mathbf{G}^{j}\right)_{q*}\right\|}. \tag{2.50}$$

Since there are $K_T = MK$ direct user channels and MQ FIS CCI channels, we obtain $n_T = MK \times MQ = M^2KQ$ such values.

2.6.3.3 MIMO Broadcast Primary Channel

Finally, when the PS is a MIMO BC with single-antenna primary receivers, we compare each one of the K_T direct user channels with each one of the MQ FIS CCI channels between each BS and each primary receiver [114]:

$$\left(\rho_{km}^{j,q}\right)^{(s)} = \frac{\left(\left(\mathbf{h}_{km}^{m}\right)^{(s)}\right)^{\dagger}\mathbf{g}_{q}^{j}}{\left\|\left(\mathbf{h}_{km}^{m}\right)^{(s)}\right\|\left\|\mathbf{g}_{q}^{j}\right\|}, \tag{2.51}$$

Again, we obtain $n_T = MK \times MQ = M^2KQ$ such values.

2.6.3.4 User Selection Rule

The sum of the inter-system correlation values is denoted as $\rho^{(s)}$, so that the average is $\left(\overline{\rho}\right)^{(s)} = \rho^{(s)}/n_T$. The scheduler selects as the set of active users the candidate user set with the minimum inter-system correlation [114]:

$$\mathcal{U}^{*} = \arg\min_{\mathcal{U}^{(s)} \in \Omega} \left(\overline{\rho}\right)^{(s)}. \tag{2.52}$$

This method is called inter-system correlation-aware scheduling. Note that the computation of the correlation values is performed over the whole search space of S candidate user sets, but *the precoders and power levels are computed only for the active set of users*.

2.7 Dynamic Cell Clustering

Now, let us consider a cellular network with M_T cells. In each cell a BS with N antennas and K single-antenna MSs are located. The cells are grouped into

M_C clusters of $M = M_T/M_C$ cells each[1]. Each BS serves its K users (which have been selected via the RSS or correlation-aware scheme) on a single time-frequency resource, but the BSs in a cluster coordinate their transmissions to mitigate the ICI. We propose a simple dynamic cell-clustering method with non-overlapping cooperation clusters, which is described next.

1. Define the total number of cells M_T, the number of cells in a cluster M, and the number of clusters $M_C = M_T/M$.
2. Define the direct neighbors of each cell C_m, $C_m^1,\ldots,C_m^L$, $m = 1,\ldots,M_T$, $L \geq M-1$.
3. Initialize the set of cells that already belong to a cluster as the empty set: $\mathcal{F}_0 = \varnothing$.
4. Initialize the set of cells that do not belong in a cooperation cluster as the set of all cells (using only their indexes): $\mathcal{V}_0 = \{1,\ldots,M_T\}$.
5. Select a random cell C_m.
6. Define the possible sets of M cells formed by the randomly selected cell and its direct neighbors, $\mathcal{S}_m^{(1)},\ldots,\mathcal{S}_m^{(\Xi)}$, where $\Xi = \binom{L}{M-1}$.
7. Compute the sum-SE achieved for each one of these sets $R_m^{(\xi)}$, $\xi = 1,\ldots,\Xi$.
8. Select as cooperation cluster $\mathcal{C}_1$ the set of M cells $\left(\mathcal{S}_m^{(\xi)}\right)^*$ that corresponds to the maximum sum-SE $\left(R_m^{(\xi)}\right)^*$:

$$\left(\mathcal{S}_m^{(\xi)}\right)^* = \arg\max_{\mathcal{S}_m^{(\xi)}} R_m^{(\xi)}. \tag{2.53}$$

9. Update the set of cells that already belong to a cluster: $\mathcal{F}_1 = \mathcal{C}_1$.
10. Update the set of cells that do not belong to a cluster accordingly: $\mathcal{V}_1 = \mathcal{V}_0 \setminus \mathcal{F}_1$.
11. Select another random cell and repeat the procedure, until the formation of $\mathcal{C}_{M_C}$.

After forming the cooperation clusters, coordinated precoding and power allocation is applied on a per-cluster level, as described in the previous sections. Note that in such a multi-cluster setup, we have to take into account the OOC interference from the direct neighbors that belong to different cooperation clusters in the computation of the sum-SE of each cluster, which is an additive interference component in the denominator of the SINR of each user.

[1] M_T is an integer multiple of M_C.

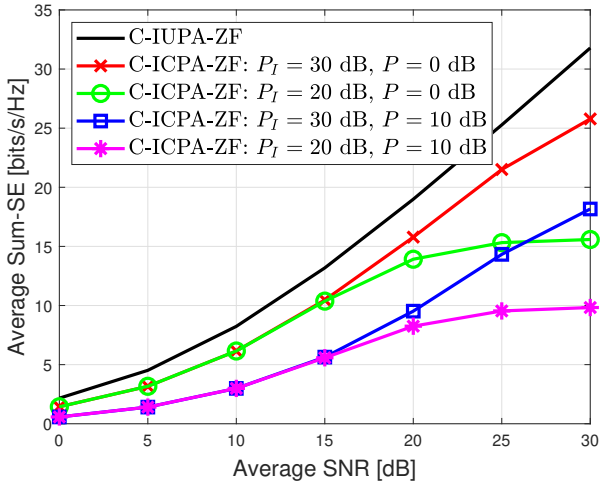
2.8 Performance Evaluation

In this section, we evaluate the proposed RA methods via numerical simulations. Initially, we consider a single-cluster cellular network that is comprised by $M = 2$ cells and is collocated with a SISO PS. A BS with $N = 4$ antennas and $K = 2$ active single-antenna MSs are located in each cell. We neglect user scheduling (i.e., the active MSs are arbitrary) and the effect of large-scale fading (i.e., we consider standard i.i.d. Rayleigh fading channels). We are interested in the average bandwidth-normalized SR (i.e., the average sum-SE) of the SS $\bar{R}$ that is achieved after 1000 simulation runs (i.e., the expected value of the bandwidth-normalized SR over 1000 channel realizations) vs. the average receive SNR $\bar{\gamma}$. The latter equals the total transmit power of each BS P_T, since both the channels and the noise have unit variance. We let P_T to vary as $P_T = \{0, \ldots, 30\}$ dB.

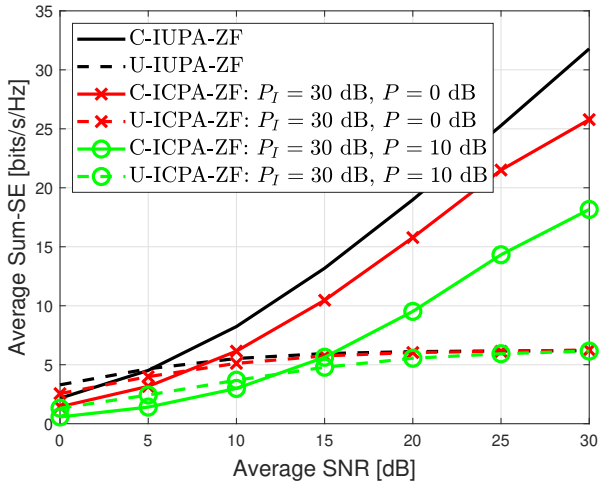
In the first test, we evaluate the performance of coordinated ZF precoding with C-ICPA (denoted as C-ICPA-ZF) for $P_I = \{30 \text{ dB}, 20 \text{ dB}\}$ and $P = \{0 \text{ dB}, 10 \text{ dB}\}$. We also plot the average SR of C-IUPA-ZF. We see in Fig. 2.2a that the average sum-SE of C-IUPA-ZF increases with the average receive SNR, as expected. The same stands for C-ICPA-ZF, although there is a performance loss due to the additional requirement of meeting the IPC as well as because of the existence of uncoordinated RIS CCI. C-ICPA-ZF performs close to C-IUPA-ZF for relaxed IPC and small P (e.g., for $(P_I, P) = (30 \text{ dB}, 0 \text{ dB})$, C-ICPA-ZF is about 3 dB worse than C-IUPA-ZF). However, its average sum-SE is significantly reduced for large values of P and floors for more hard IPC (e.g., for $P_I = 20$ dB, the capacity flooring starts to become noticeable for $\bar{\gamma} = 15$ dB). Next, we compare the performance of C-ICPA-ZF for (P_I, P) equal to (30 dB, 0 dB) and (30 dB, 10 dB) with the one achieved under the same scenarios when uncoordinated ZF precoding and ICPA (denoted as U-ICPA-ZF) is used. We plot also the sum-SE that is achieved when C-IUPA-ZF and U-IUPA-ZF are utilized in an isolated cellular network. These latter methods serve as performance benchmarks. Notice that in uncoordinated PA, the allocated power levels are determined individually in each cell. Thus, (i) there is a single Lagrange multiplier ν in the corresponding WF-PA algorithm (i.e., a single SPC); and (ii) in U-ICPA the FIS CCI consists of the channel between the BS of interest and the primary receiver. We see in Fig. 2.2b that in all cases, the average sum-SE of U-ZF precoding floors quickly due to the existence of uncoordinated ICI. This method performs slightly better than C-ZF precoding only in the noise-limited low SNR regime, where the exploitation of the spatial degrees-of-freedom (DoF) to completely eliminate the ICI might be considered an "overkill" from a sum-SE perspective[2].

In Fig. 2.3a is illustrated the performance of CQA-ICPA-ZF under a scenario where $(P_I, P) = (20 \text{ dB}, 0 \text{ dB})$ for three different QoS classes, which are defined by the minimum required power levels $\tilde{P}_{mk}^{m}$ of the $K_T = 4$ users ($m \in \mathcal{M}$, $k \in \mathcal{K}$):

[2]This is the same reason why RZF performs better than ZF in the low SNR regime.



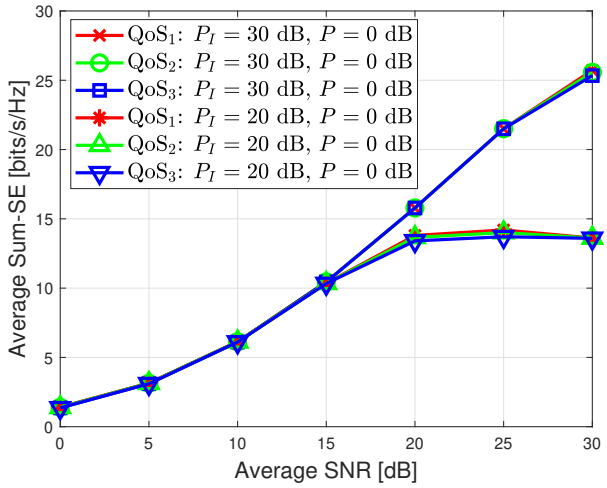
(a) C-IUPA-ZF vs. C-ICPA-ZF.



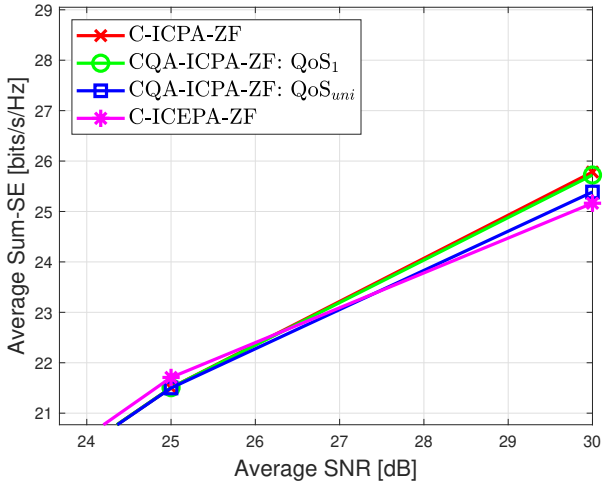
(b) C-IUPA-ZF / U-IUPA-ZF vs. C-ICPA-ZF / U-ICPA-ZF.

Figure 2.2 C-IUPA-ZF vs. C-ICPA-ZF vs. uncoordinated ZF and IUPA / ICPA.

$\text{QoS}_1 = \begin{bmatrix}0.10 & 0.10 & 0.10 & 0.10\end{bmatrix}\bar{\gamma}$, $\text{QoS}_2 = \begin{bmatrix}0.10 & 0.15 & 0.20 & 0.10\end{bmatrix}\bar{\gamma}$, and $\text{QoS}_3 = \begin{bmatrix}0.30 & 0.20 & 0.25 & 0.15\end{bmatrix}\bar{\gamma}$, listed from the less demanding to the most demanding one in terms of the minimum required rate per user. We make the assumption that when CQA-ICPA-ZF cannot meet the QoS requirements of the users, it is switched to C-ICEPA-ZF. We see that we achieve slightly better performance in the



(a) CQA-ICPA-ZF.



(b) C-ICPA-ZF vs. CQA-ICPA-ZF vs. C-ICEPA-ZF with $P_I = 30$ dB and $P = 0$ dB.

Figure 2.3 CQA-ICPA-ZF for various QoS classes and comparison with C-ICPA-ZF and C-ICEPA-ZF.

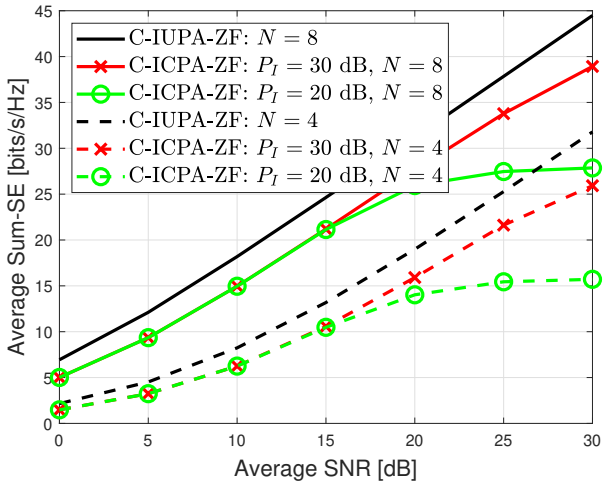
high SNR regime when the QoS requirements of the users are not that high. The reason is twofold: (i) For lower minimum rate requirements, CQA-ICPA almost resembles C-ICPA. (ii) For tight MRCs, C-ICEPA is used instead of CQA-ICPA over a large range of SNR values. In Fig. 2.3b is depicted the performance of

CQA-ICPA-ZF under the previous scenario for two different QoS classes, namely, QoS_1 and QoS_{uni}. In the latter one, the minimum required power $\tilde{P}_{mk}^{m}$ of each user takes random values between $0.10\bar{\gamma}$ and $0.30\bar{\gamma}$ that are drawn from a uniform distribution in each simulation run, to represent the variability in MRCs that is observed in practice. For comparison purposes, we present also the performance of C-ICPA-ZF and C-ICEPA-ZF. We note that the CQA-ICPA-ZF curve lies in between the C-ICPA-ZF and C-ICEPA-ZF curves in the high SNR regime. Also, for QoS_1 the performance is slightly worse than the one achieved by C-ICPA, whereas for QoS_{uni} it is slightly better than that of C-ICEPA. Notice that in Fig. 2.3b we have zoomed-in in the high SNR to show the miniscule differences in sum-SE between the different transmission strategies.

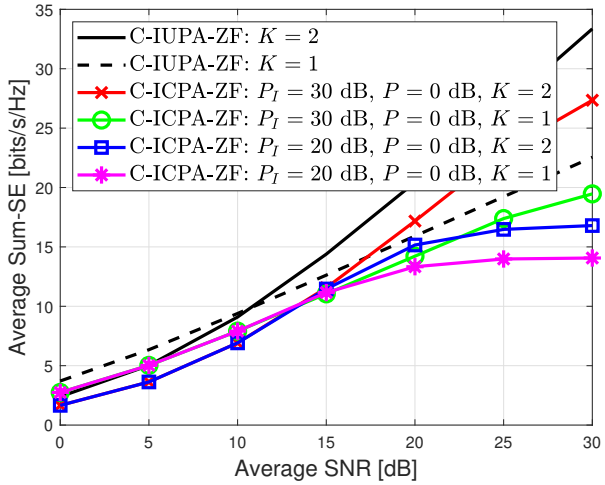
Fig. 2.4a shows the performance of C-ICPA-ZF that is achieved for (P_I, P) equal to (30 dB, 0 dB) and (20 dB, 0 dB), as well as the performance that is achieved when C-IUPA-ZF is applied in an isolated network, under two scenarios: (i) $N = 4$, as previously. (ii) $N = 8$. We note that, naturally, the performance improves substantially as the number of antennas per BS increases, thanks to the additional spatial DoF. Fig. 2.4b illustrates the performance of C-ICPA-ZF for the same (P_I, P) values as above and assuming $N = 4$, under two scenarios: (i) $K = 1$. (ii) $K = 2$, as previously. We observe that, naturally, the average sum-SE increases with the number of users per cell. However, this performance improvement is not as high as one might expected (e.g., the average sum-SE of C-ICPA-ZF at an average receive SNR of $\bar{\gamma} = 20$ dB under a scenario where $(P_I, P) = (30 \text{ dB}, 0 \text{ dB})$ and $K = 1$ is $\bar{R} = 14.22$ bps/Hz and it becomes $\bar{R} = 17.17$ bps/Hz when $K = 2$). This is due to the additional interference that is introduced into the system as well as because of the decrease in the number of spatial DoF per user as the number of users increases.

In Fig. 2.5a is shown the performance of C-ICPA when MRT, C-ZF, or C-RZF is employed under a scenario where $(P_I, P) = (30 \text{ dB}, 0 \text{ dB})$ vs. the performance of their C-IUPA counterparts, assuming $K = 2$. We notice that C-RZF outperforms C-ZF, especially in the low and moderate SNR regime, and MRT floors quickly due to the uncoordinated CCI, as expected. We also notice that the performance gap between C-IUPA-MRT and C-ICPA-MRT is very small, since the capacity saturates early even when there is no IPC. Fig. 2.5b depicts the performance of C-ICPA-RZF and C-ICPA-ZF for (30 dB, 0 dB) and (20 dB, 0 dB). We note that in all cases, the performance gets significantly degraded as P gets higher and the average sum-SE starts to floor as P_I increases, as expected.

In Fig. 2.6a and Fig. 2.6b is compared the performance of C-IUPA-PZF, CQA-IUPA-PZF for the QoS classes QoS_1 and QoS_{uni}, and C-IUEPA-PZF vs. the performance of C-ICPA-ZF, CQA-ICPA-ZF for the same QoS classes, and C-ICEPA-ZF under two scenarios where (P_I, P) is equal to (5 dB, 0 dB) or (0 dB, 0 dB), respectively. We note that for such hard IPCs, P-ZF outperforms significantly C-ZF, with the larger SE gain noticed for $P_I = 0$ dB. This is because P-ZF is not affected by P_I,



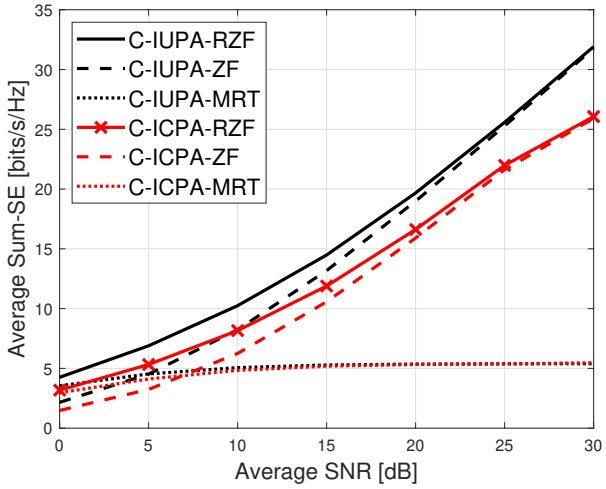
(a) C-ICPA-ZF for $N = 4$ and $N = 8$.



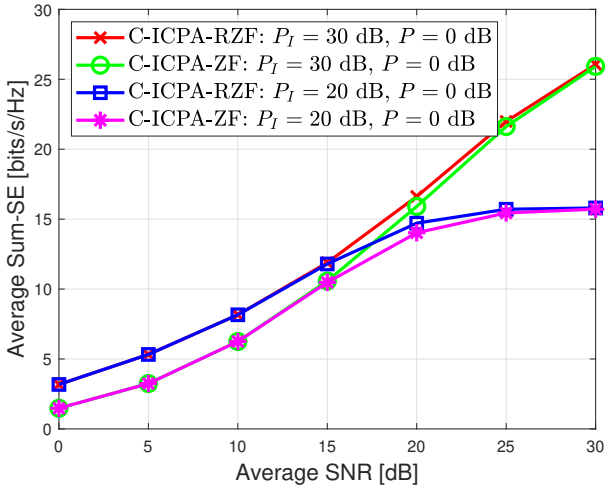
(b) C-ICPA-ZF for $K = 1$ and $K = 2$.

Figure 2.4 C-ICPA-ZF for varying number of antennas or users.

in contrast to C-ZF, but only by P. We also notice that the P-ZF variants perform almost identical in the high SNR regime, due to the uncoordinated intra-SS CCI. Furthermore, we observe that the capacity of all C-ZF variants floors due to the tight IPT. C-ICPA performs much better than its QoS-aware variants, which have also to meet the MRCs, and the inefficient C-ICEPA method. Finally, we see that



(a) C-ICPA vs. C-IUPA for MRT, C-ZF, and C-RZF with $P_I = 30$ dB, $P = 0$ dB.

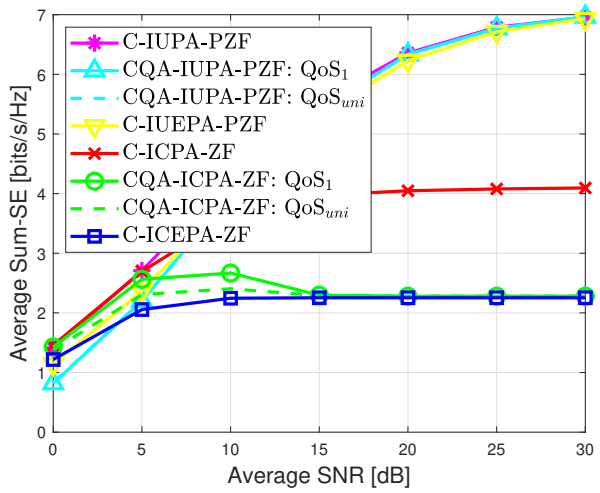


(b) C-ICPA-RZF vs. C-ICPA-ZF.

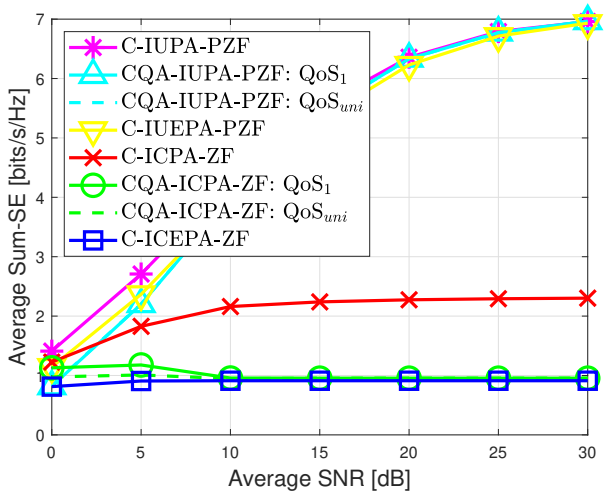
Figure 2.5 C-ICPA vs. C-IUPA for various linear precoding schemes.

C-ICPA-ZF and CQA-ICPA-ZF for QoS_1 slightly outperform C-IUPA-PZF when $P_I = 5$ dB in the noise-limited low SNR regime.

In Fig. 2.7a and Fig. 2.7b is compared the performance of C-IUPA-PZF vs. the one achieved by C-ICPA-RZF and C-ICPA-ZF for (P_I, P) equal to (5 dB, 0 dB) and



(a) C-IUPA-PZF vs. C-ICPA-ZF variants for $P_I = 5$ dB and $P = 0$ dB.



(b) C-IUPA-PZF vs. C-ICPA-ZF variants for $P_I = 0$ dB and $P = 0$ dB.

Figure 2.6 Projected ZF precoding vs. coordinated ZF precoding for different IPTs.

(0 dB, 0 dB) under two scenarios where $N = 4$ or $N = 8$, respectively. We observe that C-RZF outperforms C-ZF in the low SNR regime, but as the average SNR grows its performance starts to floor and becomes almost identical with that of C-ZF in the high SNR regime due to the hard IPC. Moreover, we notice that C-RZF and C-ZF perform slightly better than P-ZF in the low SNR regime. For moderate

and high SNR values, though, P-ZF significantly outperforms them, especially for more tight IPCs. Finally, the performance improves (and the gap between P-ZF and C-RZF / C-ZF increases) significantly when $N = 8$. In this scenario, C-ZF performs very close to C-RZF across the whole average receive SNR range, since the increased number of spatial DoF improves its performance for low SNR.

In summary: C-ICPA with ZF precoding approaches its C-IUPA counterpart for relaxed IPTs and small RIS CCI. CQA-ICPA achieves a performance in between of that of C-ICPA and ICEPA. The performance is slightly better for more relaxed MRCs. The performance of these schemes improves for more BS antennas or when RZF precoding is applied instead of ZF precoding. Finally, the use of C-IUPA variants with projected ZF precoding improves substantially the performance for hard IPTs.

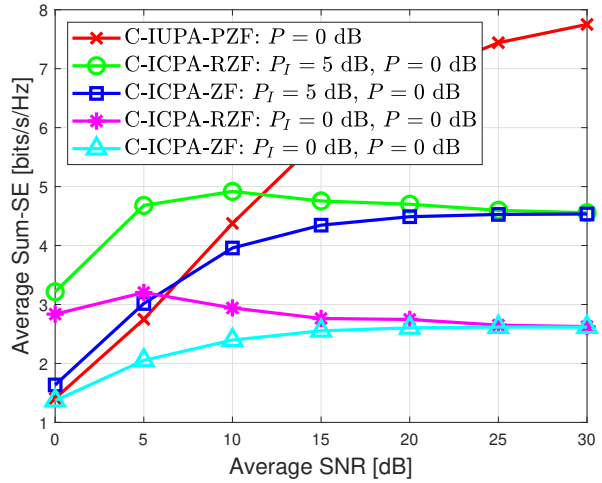
In Fig. 2.8a is compared the performance of C-ICPA-ZF for (P_I, P) equal to (30 dB, 0 dB) and (20 dB, 0 dB) against the performance of C-IUPA-ZF under three PS setups: (i) A SISO primary link. (ii) A 2×2 MIMO primary link. (iii) A 4×4 MIMO primary link. We note that the performance is worse when the PS is a MIMO link—and degrades more, the more antennas the primary receiver has or / and the harder the IPC is. We observe that for $P_I = 20$ dB, the SR starts flooring at an average SNR of 18 dB or 12 dB when the PS is a 2×2 or a 4×4 MIMO link, respectively. In Fig. 2.8b, (P_I, P) is fixed to (30 dB, 0 dB) and the previous test is repeated for both $N = 4$ and $N = 8$. We note that the performance improves in all cases when $N = 8$, as expected, but the performance gap between the SISO case and the MIMO ones remains essentially the same.

In Fig. 2.9 is illustrated the performance of C-IUPA-PZF and CQA-IUPA-PZF for QOS_1 and $P = 0$ dB when the PS is a SISO or a 2×2 MIMO link. We note that the performance is much better when the PS is a SISO link. We also notice that the performance of the QoS-aware variants converges to that of the corresponding QoS-agnostic ones for high SNR.

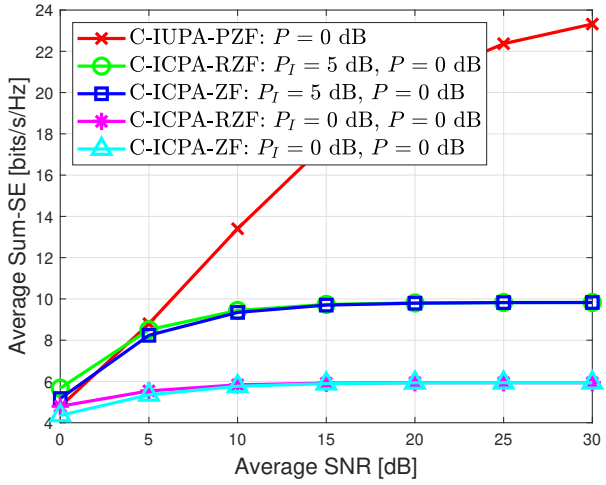
Fig. 2.10 shows the performance of C-ICPA-ZF for $P_I = 30$ dB and $P = 0$ dB when the BS is either a SISO link or a MIMO BC with $L = 2$ and $Q = 2$ or with $L = 4$ and $Q = 4$, i.e., a (2,(2,1)) or a (4,(4,1)) MIMO BC. We note that the performance falls as more primary receivers are added.

In summary: The performance degrades with the number of antennas at the primary receiver or with the number of primary receivers.

In Fig. 2.11a is illustrated the performance of optimal user scheduling based on exhaustive search vs. the performance of reduced search space (RSS) user scheduling for a use case where 2 out of 4 users are selected in each cell. C-ICPA-ZF is utilized and the following pairs of (P_I, P) values are considered: (30 dB, 0 dB), (20 dB, 0 dB), (30 dB, 10 dB), and (20 dB, 10 dB). Note that since $O = 6$ (i.e., there are 6 combinations of $K = 2$ out of $U = 4$ users in each cell) and we have $M = 2$ cells, the total number of candidate user sets is $S = O^M = 36$. In RSS scheduling, we have set $S' = 10$, i.e., the reduced search space is less than the 1/3



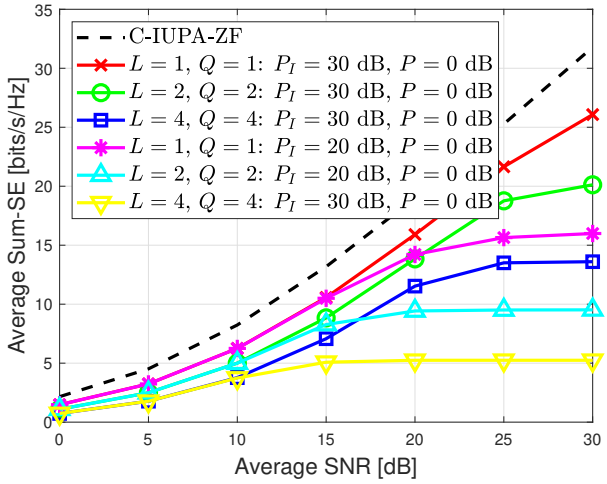
(a) C-IUPA-PZF vs. C-ICPA-RZF / C-ICPA-ZF for $N = 4$.



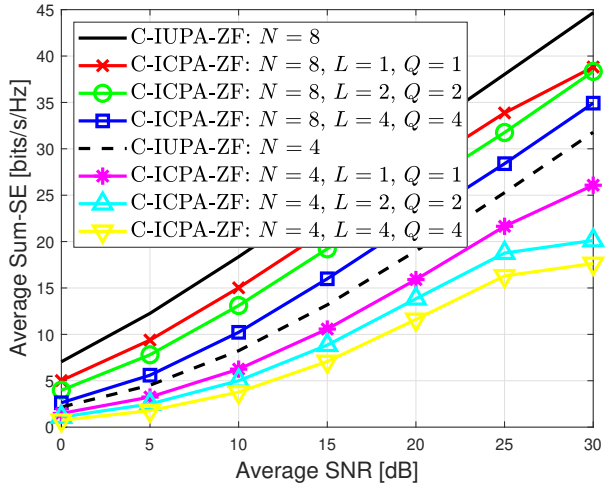
(b) C-IUPA-PZF vs. C-ICPA-RZF / C-ICPA-ZF for $N = 8$.

Figure 2.7 Projected ZF precoding vs. coordinated RZF / ZF precoding for different IPTs and numbers of antennas.

of the original search space. However, we see that the proposed scheme performs very close to the optimal one. In Fig. 2.11b we repeat the previous test under scenarios with (P_I, P) equal to (30 dB, 0 dB) or (20 dB, 0 dB) for two use cases: one where the PS is a SISO link and another where it is a 2×2 MIMO link. We see that when the PS is a MIMO link, the performance is reduced substantially—especially



(a) C-ICPA-ZF for a SISO and various MIMO PS setups.



(b) C-ICPA-ZF for a SISO and various MIMO PS setups and varying number of BS antennas with $P_I = 30$ dB and $P = 0$ dB.

Figure 2.8 C-ICPA-ZF for SISO and various MIMO PS setups.

for more stringent IPT values. In fact, already for an IPT value of $P_I = 20$ dB, we notice that the capacity starts to floor.

In Fig. 2.12a is presented the performance of RSS scheduling with $S' = 10$ vs. the one achieved for $S' = 5$ as well as vs. the performance of inter-system correlation aware user selection for P_I equal to either 30 dB or 20 dB and $P = 0$ dB,

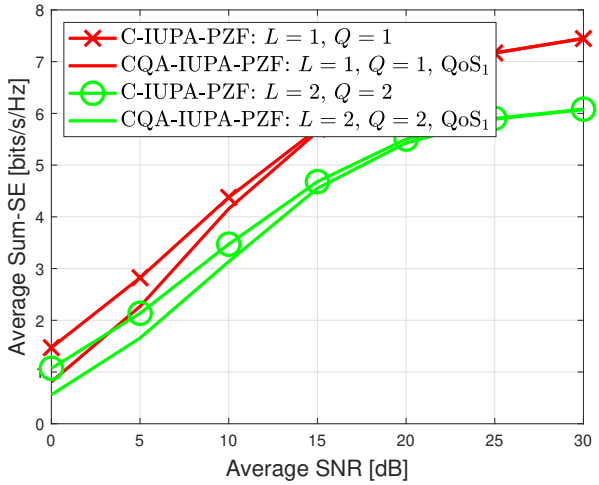


Figure 2.9 QoS-aware and QoS-agnostic projected ZF precoding for a SISO and a MIMO PS setup assuming $P = 0$ dB.

assuming the application of C-ICPA-ZF. We notice that the performance of RSS scheduling is reduced slightly when S' decreases. However, in both cases RSS performs much better than correlation-aware user scheduling. Finally, Fig. 2.12b illustrates the performance of CQA-ICPA-ZF for the QoS class QoS_1 under a scenario where $P_I = 20$ dB and $P = 0$ dB, assuming the application of RSS scheduling with $S' = 10$ or $S' = 5$ or considering the use of inter-system correlation aware user selection. We observe the same performance trends as in the C-ICPA-ZF case.

Finally, in Fig. 2.13 we compare the greedy implementation of RSS with the non-greedy one for (P_I, P) equal to (30 dB, 0 dB) and (20 dB, 0 dB) and $S' = 10$. We note that the greedy implementation results in a minor performance loss.

In summary: RSS user scheduling performs only slightly worse than optimal user selection even for moderate values of S'. The performance degrades as S' is reduced. However, RSS scheduling performs in general better than the inter-system correlation-aware scheduling scheme. The proposed greedy implementation results in a small reduction of the achieved SR.

In Fig. 2.14 is presented the performance of a multi-cluster setup where coordination is restricted within each cluster. We consider $M_T = 16$ rectangular cells with a side of 200 m divided into $M_C = 8$ clusters of $M = 2$ cells each. Each cell has $K = 2$ single-antenna MSs and a BS with $N = 4$ antennas. Cell clustering is performed dynamically as described in Section 2.7. We compare the performance with the one achieved when the cooperation clusters are fixed and their formation is based solely on the adjacency of the cells. This system-level simulation takes into account both large-scale fading and small-scale fading. The large-scale fad-

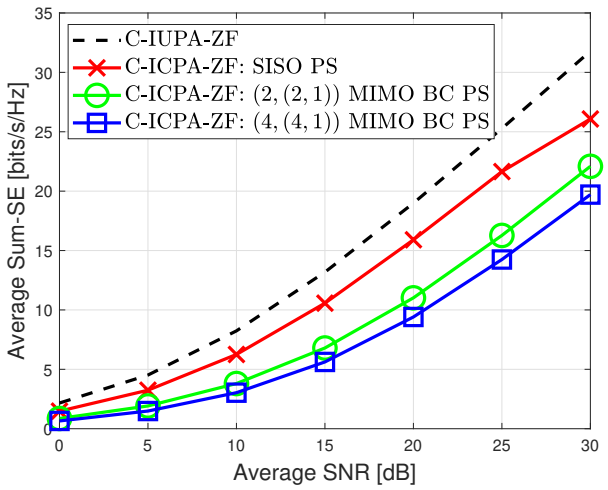


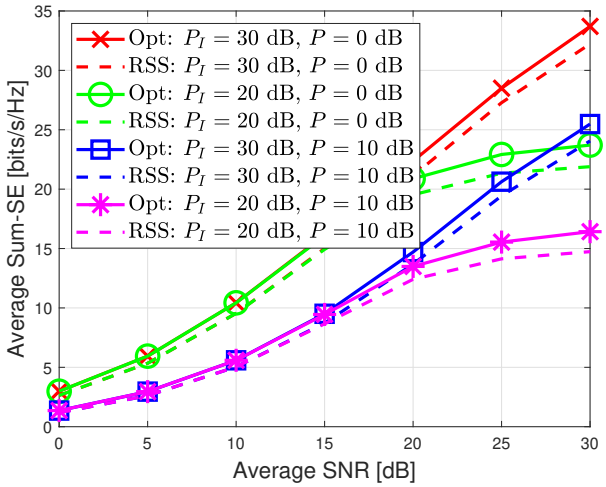
Figure 2.10 C-ICPA-ZF for a SISO and various MIMO BC PS setups with $P_I = 30$ dB and $P = 0$ dB.

ing coefficient (in dB) associated with the channel between MS_{km} and BS_j, β_{km}^j, is calculated according to the log-distance path loss with log-normal shadowing model described in Chapter 1. The parameters of the model are based on the 3GPP non-line-of-sight (NLOS) macro-cellular scenario and are summarized in Table 2.3.

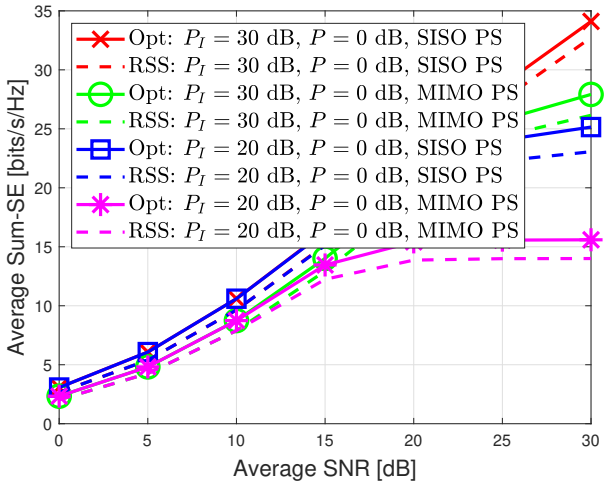
We observe in Fig. 2.14 that C-ICPA-ZF performs slightly worse than C-IUPA-ZF for relaxed IPT and low TX_{PS} transmission power. The achieved SE for both schemes is little lower in this test where we consider large-scale losses and multiple clusters. We also note that the dynamic cell clustering (DCC) scheme performs significantly better than the fixed clustering one.

2.9 Summary and Conclusions

In this work, we studied coordinated precoding, power allocation, and user selection techniques based on standard linear precoding schemes for sum-rate maximization under minimum per-user rate constraints in underlay spectrum sharing setups. Our study revealed that the use of standard linear precoding schemes and coordinated interference-aware power allocation achieves high sum-SE and makes possible QoS provisioning for relaxed IPTs. For hard IPTs, projected ZF precoding improves substantially the performance. Also, the low-complexity RSS user scheduling scheme performs closely to the optimal user selection strategy that is based on exhaustive search over the whole search space – even under a greedy im-



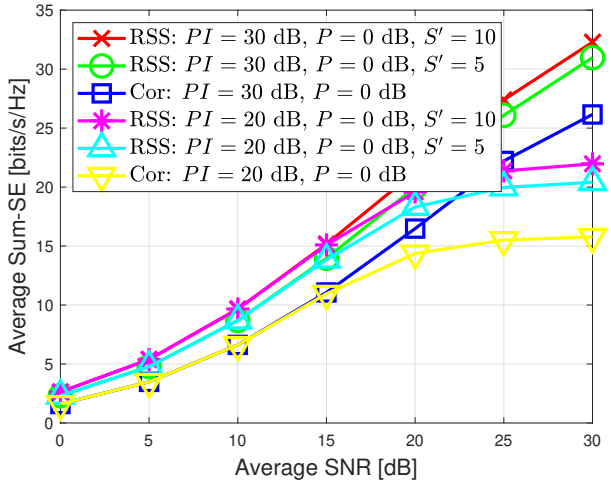
(a) Optimal user scheduling vs. RSS user scheduling.



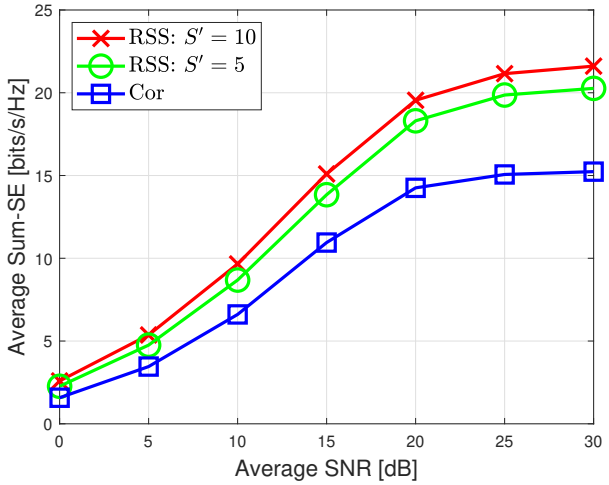
(b) Optimal user scheduling vs. RSS user scheduling for a SISO or a 2×2 MIMO PS setup.

Figure 2.11 Optimal user scheduling vs. RSS user scheduling with $S' = 10$ for various IPTs under a SISO or a MIMO PS setup.

plementation. Finally, the proposed dynamic cell clustering method outperforms fixed cell clustering.



(a) RSS scheduling with $S' = 10$ or $S' = 5$ vs. inter-system correlation-aware scheduling, assuming the application of C-ICPA-ZF.



(b) RSS scheduling with $S' = 10$ or $S' = 5$ vs. inter-system correlation-aware scheduling, assuming the application of CQA-ICPA-ZF

Figure 2.12 RSS user scheduling vs. correlation-aware user scheduling for different values of S'.

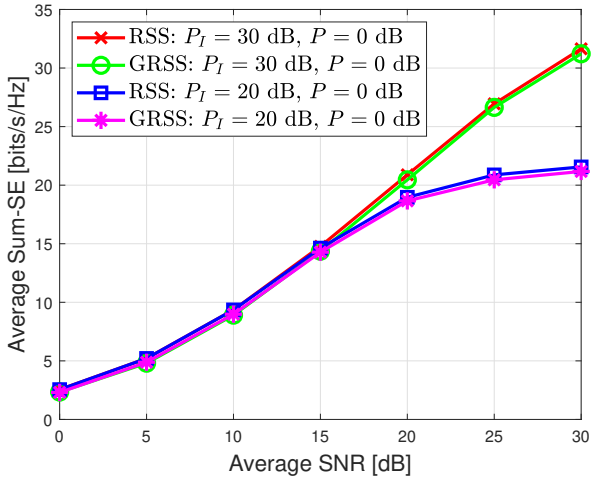


Figure 2.13 RSS vs. GRSS with $S' = 10$.

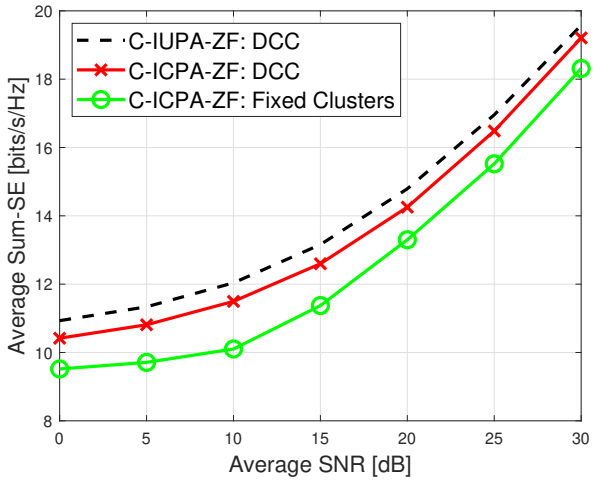


Figure 2.14 System-level performance: dynamic vs. fixed cooperation clusters.

Table 2.3 System-Level Simulation Parameters

System-Level Simulation Parameters	
Frequency	$f_c = 2$ GHz
PLE	$\zeta = 3.76$
Shadowing s.d.	$\sigma_{sf} = 10$ dB
Reference distance	$r_0 = 1$ m
Mean path loss at r_0	$\overline{L}_p(r_0) = 35.3$ dB
Type of cells	Square – BS is placed in the center
Network setup	Square Grid
Cell size	Side with length of 200 m
Total number of cells	$M_T = 16$
Number of cooperation clusters	$M_C = 8$
Number of cells per cooperation cluster	$M = 2$
Number of BS antennas	$N = 4$
Number of single-antenna MSs per cell	$K = 2$

Chapter 3

Spectrum Sharing II: Cache-Aided Joint Transmission

3.1 Introduction

The joint transmission (JT) variant of coordinated multi-point (CoMP) is rarely applied in practice, although it further improves the QoS of the cell-edge users in comparison to the coordinated precoding (CP) variant. This is because, in contrast to CoMP-CP, it requires also the sharing of user data among the cooperating BSs in addition to CSI, thus imposing a heavy burden on the mobile transport network in terms of throughput and latency requirements [5]. Mobile edge caching have been proposed as a workaround to this problem. This technology not only reduces the transport delays and offloads the network by enabling local service provisioning [88–90], but creates also JT opportunities through the exploitation of the redundancy in the stored contents [91].

3.1.1 Motivation and Related Work

An underlay spectrum sharing paradigm that makes use of cache-aided CoMP-JT should utilize interference coordination when such cache-enabled transmissions are not possible, in order to be efficient. Nevertheless, most studies consider uncoordinated transmissions in this case [105, 106]. This approach is highly suboptimal from a sum-SE maximization and interference management perspective and does not suit the underlay spectrum sharing context under consideration.

Furthermore, the applied caching scheme should be efficient. Recal that a cache server stores frequently requested content to serve subsequent user requests rather

locally than from the remote origin servers. The caching algorithm that runs on a cache server determines which content will enter or get evicted from the local storage, so that the performance is optimized w.r.t. a given metric. The main performance measure is the cache hit rate, i.e., the fraction of user requests that are served by the cache. This metric provides an indication of both the traffic savings and latency reduction associated with caching. Least recently used (LRU) constitutes the most commonly employed caching scheme, due to its simple software implementation, constant $\mathcal{O}(1)$ cache update effort per request, and ability of adapting to the temporal dynamics of the access pattern. On the other hand, this caching strategy is highly inefficient, in terms of the achieved cache hit rate. Several alternatives that significantly outperform LRU while preserving its beneficial characteristics have been studied in the literature [88, 89, 119], but not under a cache-aided CoMP-JT context.

3.1.2 Contributions

In this chapter, we present coordinated caching strategies that create JT opportunities as well as simple caching schemes with $\mathcal{O}(1)$ update effort per request that achieve higher cache hit rate than LRU. These techniques are studied in a hybrid CoMP-CP / CoMP-JT framework, where CoMP-CP takes place whenever cache-aided CoMP-JT is not possible. The coordinated resource allocation (RA) techniques described in the previous chapter are utilized to improve the performance and protect the incumbent from harmful interference.

3.1.3 Organization

The chapter is organized as follows: In Section 3.2 is presented the system setup, while Section 3.3 introduces the system model. Section 3.4 describes the power allocation schemes. The proposed caching strategies are described in Section 3.5. The simulation results are discussed in Section 3.6. Finally, Section 3.7 provides a summary of this work and presents our conclusions.

Mathematical Notation: See Chapter 2.

3.2 System Setup

The system setup presented in Chapter 2 is considered here as well, with one difference: each BS is equipped with a cache of storage capacity $C \ll F$ files, where F is the size of the content catalog. We assume files of equal size. This is because the cache storage of BSs ranges typically from several hundreds of GBs to few tens of TBs, while codecs and transport protocols divide videos and files into small segments with size of a few MBs, thus turning bin-packing into a minor issue [88, 89, 119].

Also, we assume i.i.d. Zipf distributed user requests to the files of the catalog [88, 120, 121]. Finally, we consider a SISO primary link. Note that in CoMP-JT mode, each BS serves all the users in the cluster.

3.3 Signal Model for Joint Transmission

Similar to the CoMP-CP case, the complex baseband representation of the received signal at MS_{km} when CoMP-JT is applied is given by [122, 123]:

$$y_{km} = \sum_{j=1}^{M} \sum_{l=1}^{M} \sum_{i=1}^{K} \left(\mathbf{h}_{km}^{j}\right)^{\dagger} \mathbf{v}_{li}^{j} + h_{km}\sqrt{P}d + n_{km}. \tag{3.1}$$

The first term at the RHS of Eq. (3.1) can be decomposed into the sum of a data component, an intra-cell MUI component, and an ICI component as follows:

$$\breve{y}_{km} = \sum_{j=1}^{M} \left(\mathbf{h}_{km}^{j}\right)^{\dagger} \mathbf{v}_{mk}^{j} + \sum_{j=1}^{M} \sum_{\substack{i=1 \\ i\neq k}}^{K} \left(\mathbf{h}_{km}^{j}\right)^{\dagger} \mathbf{v}_{mi}^{j} + \sum_{j=1}^{M} \sum_{\substack{l=1 \\ l\neq m}}^{M} \sum_{i=1}^{K} \left(\mathbf{h}_{km}^{j}\right)^{\dagger} \mathbf{v}_{li}^{j}. \tag{3.2}$$

The complex baseband representation of the received signal at $\mathrm{RX}_{\mathrm{PS}}$ is given by:

$$y = g\sqrt{P}d + \sum_{m=1}^{M} \sum_{l=1}^{M} \sum_{k=1}^{K} \left(\mathbf{g}_m\right)^{\dagger} \mathbf{v}_{lk}^{m} + z. \tag{3.3}$$

3.4 Power Allocation

The use of coordinated interference-constrained equal power allocation (C-ICEPA) is considered for both CoMP-CP and CoMP-JT. Here is presented this PA technique for both CoMP variants [122, 123]:

$$P_c = \begin{cases} \min\left(\frac{P_T}{K}, \frac{P_I}{\sum_{m=1}^{M}\sum_{k=1}^{K} \alpha_{mk}^{m}}\right) & \text{CoMP-CP} \\ \min\left(\frac{P_T}{MK}, \frac{P_I}{\sum_{j=1}^{M}\sum_{m=1}^{M}\sum_{k=1}^{K} \alpha_{mk}^{j}}\right) & \text{CoMP-JT} \end{cases}. \tag{3.4}$$

Note that when the IPC is inactive, C-ICEPA is reduced to conventional C-IUEPA, i.e., $P_c = P_T/K$ for CoMP-CP or $P_c = P_T/MK$ for CoMP-JT.

3.5 Coordinated Caching

3.5.1 Zipf's Law

Several measurements of the patterns of requests for content on the Internet revealed that the user behavior is governed by Zipf's law, so that a small subset of popular objects (e.g., videos, files, web site pages, etc.) attracts the main portion of the user requests [119]. In particular, a Zipf distribution associated with a finite set $\mathcal{F}$ of F objects O_r attributes request probabilities $p(r)$ to these objects corresponding to their popularity rank $r = 1,\dots,F$, with $p(1) \geq \cdots \geq p(F)$, according to the following relation [119]:

$$p(r) = Ar^{-\beta} = \frac{r^{-\beta}}{\sum_{r=1}^{F} r^{-\beta}}, \quad A, \beta > 0, \tag{3.5a}$$

$$\sum_{r=1}^{F} p(r) = 1 \Rightarrow A = p(1) = \frac{1}{\sum_{r=1}^{F} r^{-\beta}}, \tag{3.5b}$$

where A is a normalization constant and β is a shaping factor that determines the skewness of the distribution. Typically, $\beta \in \{0.5, 1\}$, as confirmed by measurement studies considering YouTube videos, IPTV platforms, web sites, and P2P file transfer systems [120, 121, 124, 125].

The high concentration of user requests to a small number of popular objects implies that even relatively small caches can be quite efficient, in terms of the achieved cache hit rate, provided that the applied caching strategy stores the most popular objects in the cache [119].

3.5.2 Content Popularity Dynamics

In Section 3.5.1, we have implicitly assumed a stream of i.i.d. requests to the objects, so that a request refers to an object O_r with a constant probability $p(r)$. Under this independent reference model (IRM) [126], the optimum hit rate of a cache with storage capacity sufficient to hold $C \ll F$ objects equals the sum of the access probabilities of the top C objects in terms of popularity ranking, i.e., $h_{\text{opt}}^{C} = \sum_{r=1}^{C} p(r)$.

In practice, though, the popularity of the objects changes over time and new objects enter the content catalog. Nevertheless, the popularity dynamics is in general relatively low (i.e., rank changes take place in the time scale of hours or days and affect only a small subset of the objects each time [120, 125]). For such slowly varying pattern, the hit rates achieved by caches that serve a large user population and handle hundreds of thousands of requests per day are close to the ones computed under IRM conditions for Zipf distributed requests [119]. Thus, Eq. (3.5) represents a simple, yet valid approximation of content access patterns.

Therefore, we can approach the optimum hit rate in practice if we hold in the cache the most popular objects over long timeframes [88, 89, 119]. On the other hand, a practical caching scheme should be able to react to the popularity changes observed in realistic scenarios by replacing formerly popular cached objects with new ones that became "hot" in a recent timeframe, in order to avoid the pollution of the cache storage by outdated objects.

3.5.3 Caching Schemes

A caching scheme assigns values to the objects either explicitly according to a score function or implicitly via a ranking method, in order to determine which objects to store in or drop from the local storage. It is typically implemented in software as some type of list with stored objects. The computational complexity of the cache storage lookup, cache update, and object insertion / replacement operations is of major importance in practical implementations. Several caching schemes whose goal is to maximize the cache hit rate have been studied in the literature.

LRU and LFU Least recently used (LRU) ranks the objects according to their time-of-last-access and stores in a cache of size C the C most recently referenced objects, sorted in decreasing request recency order. LRU is the most widely adopted caching scheme due to its simple implementation, constant $\mathcal{O}(1)$ effort per request for putting the requested object on the top of the cache stack, and ability to adapt to the access pattern dynamics, since it promotes the caching of recently "active" objects [88, 123]. On the other hand, this caching scheme is highly inefficient, as it has been shown analytically as well as through numerical simulations and trace-based measurement studies, because it does not take into account object popularity in the caching and replacement decisions. In fact, the absence of request count statistics leads often to the pollution of LRU caches by objects that are referenced only once, which degrades the caching efficiency [88, 89, 119]. Moreover, LRU presents a high rate of loading objects into the cache, since in each cache miss the requested object is transferred to the cache storage. The frequent downloading of external objects increases the processing load, the latency, and the network traffic.

Least frequently used (LFU), on the other hand, counts the number of past requests to each object and holds in a cache of size C the C most frequently referenced objects, sorted in decreasing request frequency order. Typically, LRU is used as a tie-breaker between objects of the same value. LFU achieves the optimum hit rate under IRM conditions, since its request count statistics converge over time and reflect the popularity ranking of the objects. Also, LFU allows the caching of requested objects only when their request count is higher than (or at least equal to in some implementations) the request count of the least frequently referenced cached object, thus reducing the loading rate of external objects into the cache. However, this caching strategy is rarely used in practical applications,

since its unlimited request statistics leads to pollution of the cache by currently irrelevant objects that are maintained in the local storage over long timeframes due to their high request count and influece the caching and replacement decisions. Yet, LFU serves as a benchmark under the IRM. Another reason why LFU has not been preferred in practice is because the conventional implementation of this caching scheme presents an $\mathcal{O}(C)$ insertion, replacement, and update complexity for maintaining a perfectly sorted (w.r.t. the request count of the objects) cache list of size C. Nevertheless, we should note that there have been proposed also implementations of LFU with $\mathcal{O}(1)$ effort per request (which, however, require at least twice the time needed by LRU to perform a cache update) [127].

Design Criteria Several alternatives to these standard caching methods have been proposed in the literature. Our focus is on caching schemes that meet the following criteria [119, 123]:

1. They have simple implementation and present constant $\mathcal{O}(1)$ effort per request.
2. They approach the optimum LFU hit rate under IRM conditions.
3. They react to the dynamically changing popularity of the objects.
4. They implement some admission control and replacement policy that reduces the rate of loading external objects into the cache.
5. They provide the flexibility to consider other performance metrics than the cache hit rate.

Typically, such caching strategies inspect access statistics of past requests to extract information about the frequency and recency of requests to objects.

WLFU and WLFU-NE Window LFU (WLFU) [88] restricts the LFU principle in a sliding window (SW) of W requests, which acts as an aging mechanism that prevents cache pollution with objects of decreasing relevance [119]. The window size determines the reach of the statistics in the past and, thus, represents a single adaptation parameter for balancing the impact of request frequency and recency information on caching and replacement decisions. WLFU resembles LRU for small window sizes and approaches LFU as the window size increases. We can further simplify this caching method by performing insertion of objects always from the beginning of the cache list (i.e., in an LRU-like fashion) and by considering simple cache updates that involve the comparison of the scores of two objects, that is, the requested object and its neighbor from left in the cache list upon a cache hit (since then the score of the requested object increases by one) or of a cached object whose request dropped from the window and its neighbor from right in the

cache list (since the score of this object decreases by one). This WLFU with neighbor (position) exchange (WLFU-NE) cache updates scheme starts with a cache list that is partially sorted w.r.t. the objects' scores and over time results in a perfectly sorted cache list via the aforementioned simple cache updates [89].

SG-LRU and SG-C By using LRU-type updates instead of WLFU-NE updates, we get score-gated LRU (SG-LRU) [119]: a caching scheme that combines the LRU principle for simple implementation and fast updates with a score-gate function (here, WLFU) for avoiding the frequent loading of objects in the cache and storing the most popular objects in a recent timeframe. Fig. 3.1 depicts the operation of this proposed caching scheme with the help of an example. SG-LRU runs faster than LRU, since it avoids the frequent updates caused by cache misses. Moreover, it replaces over time lower ranked objects with higher ranked ones in the cache and approximates closely WLFU / WLFU-NE. If we omit the LRU cache structure (i.e., if upon a cache hit we simply update the score of the requested object but not its position in the cache list) and compare in each user request the score of the requested object with the score of a random cached object that is determined by a corresponding pointer ("clock hand") that cycles through the cache list, then we will obtain an even simpler score-gated clock (SG-C) scheme [127]. SG-C runs faster than SG-LRU due to the fact that the LRU updates caused by cache hits are relatively time-consuming operations, in contrast to the LRU updates caused by cache misses. Notice that SG-LRU and SG-C provide the flexibility to use an arbitrary scoring function for ranking the objects and, therefore, can optimize the performance w.r.t. any criterion. Thus, these caching strategies meet all the design criteria mentioned previously. This is in contrast to WLFU and WLFU-NE, where the scoring function defines also the cache structure / caching principle. Hence, these caching schemes do not meet the design criterion (5).

3.5.4 C3RE Caching

In this work, we propose a coordinated content caching with redundancy enhancement (C3RE) method [122, 123], where upon a local cache miss the target cache downloads the requested object from another cache in the cluster if possible (global cache hit), thus leaving the fetching of this object from the origin server as a last resort (global cache miss). Upon a global cache hit, the remote cache may update only its window (cooperation variant II.A) or both its window and local storage (cooperation variant II.B) or none of them (cooperation variant I), assuming that WLFU, WLFU-NE, or SG-LRU is utilized. For SG-C, we consider only variants I and II.B (simply referred to as II). When LRU is applied, the remote cache may (variant II) or may not (variant I) update its local storage upon a global cache hit. On the other hand, whenever a file enters the target cache, the corresponding BS updates both the window (if there is any) and local storage of its cache.

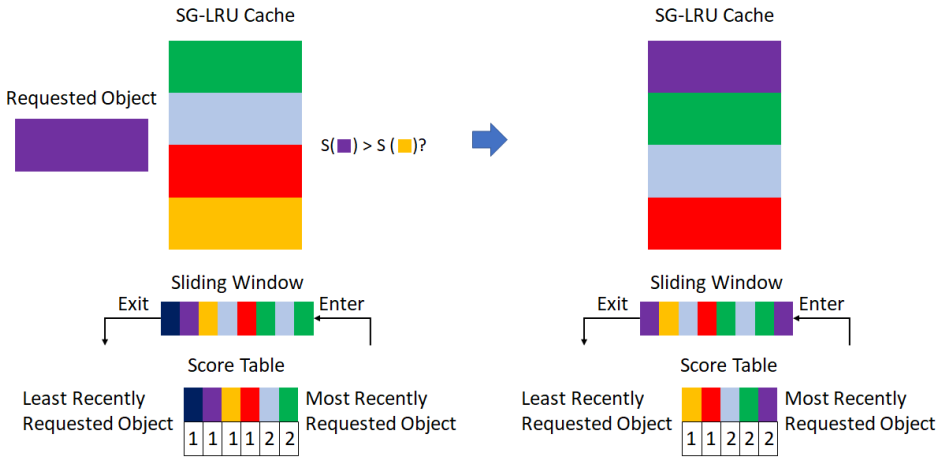


Figure 3.1 Replacement operation of an SG-LRU cache with size $C = 4$ that utilizes the WLFU score function with a sliding window of size $W = 8$.

Cooperative caching reduces the latency and traffic associated with the downloading of objects from the origin server upon local cache misses at the expense of the cooperation overhead to fetch the requested object from a remote cache. Also, the aforementioned caching variants create different levels of content redundancy across the cache servers, which can be exploited towards JT transmissions.

3.6 Performance Evaluation

3.6.1 Cache Hit Rates

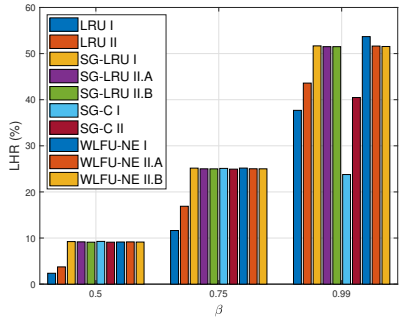
Our main motivation for applying caching has been the facilitation of CoMP-JT. However, an efficient caching strategy should also reduce the network traffic and delays by storing popular content. In this section, we study the performance of the considered C3RE variants when LRU, SG-LRU, SG-C, and WLFU-NE is applied, in terms of the average local, global, and total cache hit rate (LHR, GHR, and THR, respectively) that is achieved after $N_{\text{sim}} = 1{,}000$ simulation runs. We assume a cooperation cluster consisting of $M = 2$ cells, with $K = 1$ active user per cell that is selected in each scheduling slot from a large user population via some user selection algorithm. We also assume initially a content catalog with a size of $F = 10{,}000$ files, $N_c = 2$ cache servers (one for each BS) with a storage capacity of $C = 100$ files each (i.e., equal to the 1% of the catalog size), $N_r = 1{,}000{,}000$ user requests addressed to each cache server in every simulation run, and a window with a size of $W = 100{,}000$ requests (i.e., equal to the 10% of the requests). Note that in each simulation run, we ignore the results of the first 25% of the requests,

to exclude the cache filling phase and transient behavior from the steady-state performance evaluation.

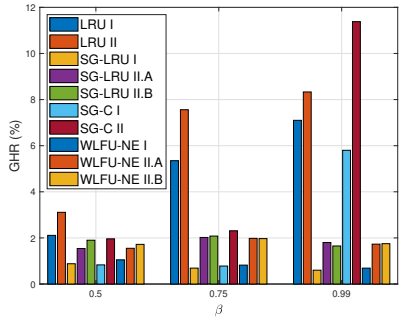
Use Case (1): We vary the shape parameter of the Zipf distribution as $\beta = \{0.5; 0.75; 0.99\}$. The hit rates are shown in Fig. 3.2 [123].

- LHR: *The LHR of all caching strategies improves as β increases*, as expected. For LRU, this is attributed to the fact that as the popular objects become "hotter", they are typically requested more often in shorter time intervals and, therefore, enter more frequently the LRU cache. The statistics-based caching schemes achieve similar LHR across the considered range of β, except for the SG-C variants whose LHR degrades for high β due to the random selection of the "least valuable" cached object, which does not let this strategy to exploit the high unbalance of user requests in favor of popular objects in this scenario. Naturally, *the statistics-based caching schemes outperform significantly LRU*, due to the exploitation of request count information in the caching and replacement decisions. We should also mention that LRU II, where the remote cache is allowed to update its local storage upon a global cache hit, achieves higher LHR than LRU I across the considered range of β. This is due to the fact that such remote cache updates provide an indirect and limited, yet useful indication of global statistics, whose exploitation leads to higher LHR for the remote cache. On the other hand, we don't notice major performance differences between the variants of the statistics-based caching schemes.

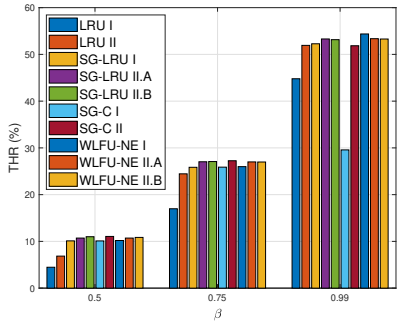
- GHR: *LRU outperforms significantly the other caching schemes, especially for moderate and large values of β*, with the exception of SG-C II which performs much better than LRU for $\beta = 0.99$. The superiority of LRU against the statistics-based strategies is explained by the fact that the remote cache acts as a second-level cache that deals with requests that have been filtered by the target cache and, therefore, do not follow a Zipf distribution. Similarly, the random selection of cached objects whose score controls whether an insertion of an external object into the cache will take place or not, makes SG-C to perform similar to LRU—or even better than LRU when the remote cache is allowed to update its scores. Furthermore, we see that the GHR of LRU improves as β gets larger, in contrast to the behavior of the remaining caching strategies. This phenomenon is attributed to the filtering of the requests from the target cache, which reduces the influence of content popularity in caching and replacement decisions. Another interesting observation is that the corresponding variants of the statistics-based caching schemes perform close to each other, as well as that *in the majority of cases the exploitation of global statistics improves the GHR*.



(a) Local hit rates.



(b) Global hit rates.



(c) Total hit rates.

Figure 3.2 Cache hit rates for varying Zipf shape parameter β.

- THR: *The THR of all caching schemes improves as β increases*. Furthermore, we see that *the statistics-based caching techniques outperform only slightly LRU*, with their performance gain over LRU being a little bit higher for larger values of

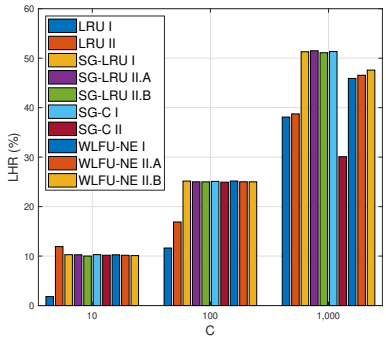
β. This is because their LHR gains are partially compensated by the GHR gains of LRU. Also, we observe that these statistics-based strategies perform close to each other, with the notable exception of SG-C I whose THR drops significantly when $\beta = 0.99$ (in comparison to other similar strategies).

Use Case (2): We set $\beta = 0.75$ and vary the cache size as $C = \{10;100;1,000\}$ (i.e., as $\{0.1\%;1\%;10\%\}$ of the catalog size). We should mention that a cache size equal to the 10% of the catalog size is rather unrealistic and it has been added here only as a means to study the cache behavior in extreme conditions. The hit rates are shown in Fig. 3.3.

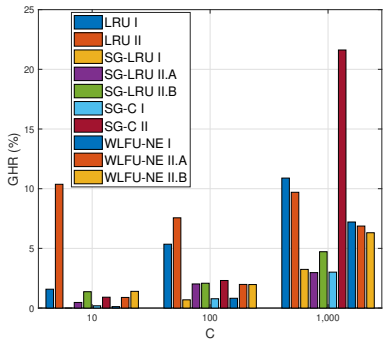
- LHR: *The LHR of all caching strategies grows with the cache size*, as expected. The statistics-based caching schemes perform close to each other, with the exception of the case for $C = 1,000$ where the performance of the WLFU-NE variants and of SG-C II is a little worse and considerably worse, respectively, than the performance of the SG-LRU variants and SG-C I. The small performance degradation of WLFU-NE for large cache size is caused mainly by the fact that the successive neighbor exchange updates take a long time to produce a sorted cache list w.r.t. the score of the cached objects. On the other hand, the significant performance degradation of SG-C II for large caches is attributed to the fact that such caches store along with the few very popular objects a large number of not so popular objects. In this case, the comparison of an external object's score with the score of a pretty-much randomly selected cached object is highly inefficient. Also, for the same reason the update of the remote cache's scores upon a global cache hit further decreases its LHR. *The statistics-based caching methods outperform LRU, except for small caches where LRU II (which constantly outperforms LRU I) performs better*. This is because the influence of request count statistics on the caching efficiency is smaller for small caches.

- GHR: *LRU outperforms significantly the statistics-based caching schemes, especially for small caches*, with the exception of SG-C II for $C = 1,000$. Also, *variants II perform in general better than variants I*, with few exceptions when $C = 1,000$.

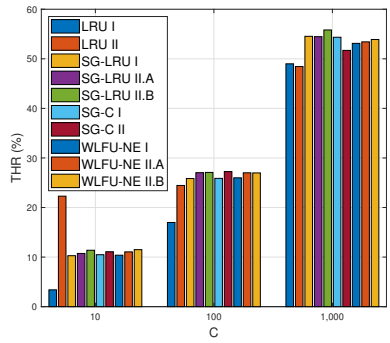
- THR: *The THR of all caching schemes improves as C increases (less aggressively for LRU when C is small). LRU II outperforms the statistics-based caching strategies only for* $C = 10$. The statistics-based caching variants perform similar to each other. In general, *variants II present a little higher THR than variants I*, with the exception of SG-C for the case where $C = 1,000$ due to the fact that its high GHR is compensated by its low LHR. SG-LRU II.B performs slightly better than the other statistics-based caching schemes across the whole range of C.



(a) Local hit rates.

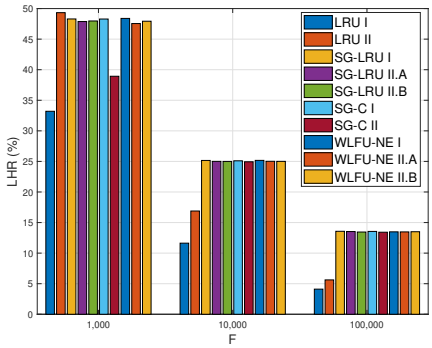


(b) Global hit rates.

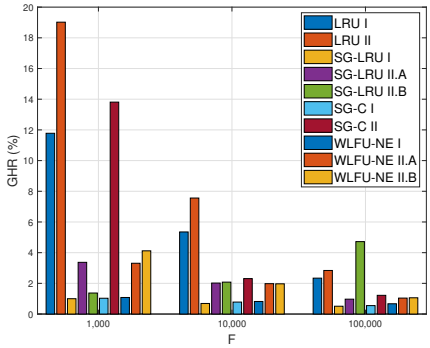


(c) Total hit rates.

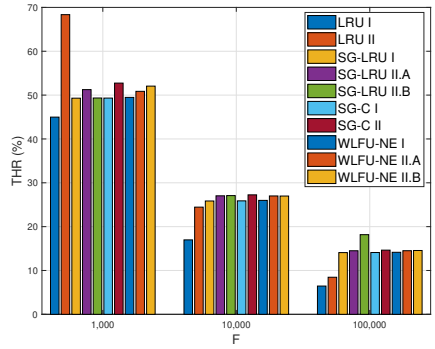
Figure 3.3 Cache hit rates for varying cache size C.



(a) Local hit rates.



(b) Global hit rates.



(c) Total hit rates.

Figure 3.4 Cache hit rates for varying catalog size F.

Use Case (3): We fix $C = 100$ and vary the catalog size, so that the cache size C is the $\{10\%;1\%;0.1\%\}$ of the catalog size, as in Use Case (2), i.e., $F = \{1000;\ 10{,}000;\ 100{,}000\}$. The hit rates are shown in Fig. 3.4. We observe similar behavior with the one illustrated in Fig. 3.3 for the corresponding catalog size / cache size ratios, with a few exceptions that we present here.

- LHR: The performance of WLFU-NE does not present any degradations, since the cache size is moderate. Also, LRU II presents a small performance gain over the statistics-based caching strategies for small catalog size.
- GHR: SG-LRU II.B performs better than LRU for large catalog size / cache size ratio. Moreover, SG-LRU II.A presents a small performance gain over SG-LRU II.B for small catalog size, whereas for large catalog size we see the opposite phenomenon.
- THR: LRU II outperforms significantly the statistics-based caching schemes for small catalog size. SG-C II and SG-LRU II.B achieve the highest THR among the statistics-based caching methods for small and large catalog size, respectively.

Use Case (4): We set $F = 10{,}000$ and vary the number of user requests as $N_r = \{100{,}000;500{,}000;1{,}000{,}000\}$. The hit rates are depicted in Fig. 3.5. We note that SG-LRU and LRU I have almost reached their steady-state LHR already after $100{,}000$ requests, while SG-C and WLFU-NE needed $500{,}000$ requests due to the random selection of the least valuable cached object and the successive cache updates based on NE, respectively. Also, we see that LRU II converged after $1{,}000{,}000$ requests due to the filtering of the requests seen by the remote cache. The same picture holds true for the GHR. Regarding the THR, we see that LRU I and the statistics-based caching strategies have approached their steady-state performance after $100{,}000$ requests, whereas LRU II needed $1{,}000{,}000$ requests.

Use Case (5): We set the number of requests to $N_r = 1{,}000{,}000$ and vary the window size of the statistics-based caching schemes from $W = 1$ up to $W = 1{,}000{,}000$, i.e., from the 0.0001% up to the 100% of the number of user requests. The hit rates are depicted in Fig. 3.6. We note that the LHR and THR of all caching strategies increases as W grows while the GHR decreases. This is because (i) in larger windows the request frequency information influences more the caching and replacement decisions than the request recency information, thus improving the LHR and decreasing the GHR; and (ii) the LHR gains compensate for the GHR losses. For small window size, the statistics-based caching strategies present LRU-like behavior and they start to approach their baseline performance that is achieved for $W = 100{,}000$ for moderate window sizes, with WLFU-NE having a slower convergence rate than the other caching methods. SG-LRU II.B constitutes an exception,

since its performance is relatively high even for $W = 1$. When the window size becomes practically unlimited (i.e., equal to the number of requests), the hit rates remain constant or vary slightly, except for SG-C I whose LHR and GHR drops and increases significantly, respectively.

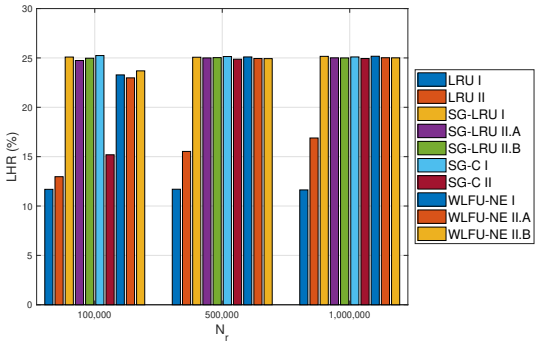
Summary: The statistics-based strategies achieve higher LHR than LRU (except for small caches), with the best performance obtained for large values of β, C, and W. SG-C constitutes an exception, since for large values of β or C or for practically infinite window size, it presents LRU-like behavior (like most statistics-based methods do for small window size). LRU, on the other hand, outperforms these caching schemes in terms of GHR. In most cases, the use of global statistics improves the GHR. We should note that the caching methods that achieve high LHR are preferable over caching schemes that achieve high GHR and have comparable THR with them, since local caching results in higher delay reduction and network traffic savings.

3.6.2 Joint Transmission Opportunities

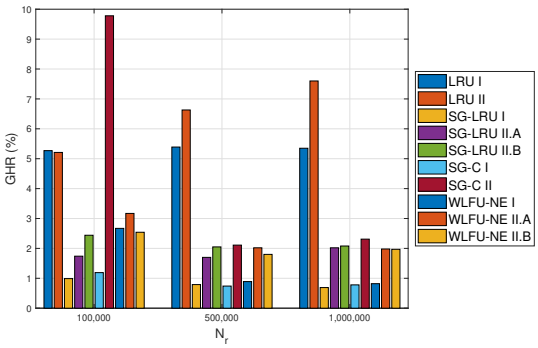
Next, we study the ability of the considered caching strategies to create JT opportunities in the aforementioned use cases. Note that if MS_{11} (the sole active user associated with BS_1) requests O_i and MS_{12} (the sole active user associated with BS_2) requests O_j (O_i, $O_j \in \mathcal{F}$, $i \neq j$), then each one of C_1 and C_2 should have stored both these objects, for cache-aided JT to take place.

In Fig. 3.7a we see that the percentage of JT opportunities grows with β for all caching schemes, as expected. This is because the caching efficiency improves for higher values of β and the considered caching methods exploit (in a bigger or smaller extent) content popularity information in the caching and replacement decisions, thus ending up often with similar cache lists. SG-C I constitutes an exception, since the performance of this caching strategy drops for $\beta = 0.99$. This is due to its LHR loss in this scenario that is caused by the random selection of the candidate for eviction cached object. Naturally, the number of JT opportunities is small for low values of β and all caching strategies perform close to each other in this scenario. For moderate and high values of β, the statistics-based caching methods outperform significantly LRU (with the exception of SG-C), thanks to the direct exploitation of request frequency information.

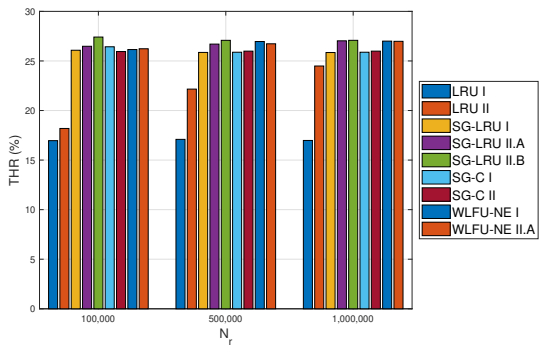
In Fig. 3.7b we note that the number of JT opportunities increases with the cache size, as expected. This is because the caching efficiency improves for larger cache sizes and the probability of finding objects that have been stored in both caches increases for larger cache lists. Again, SG-C II represents an exception, since its performance drops for $C = 1,000$. The statistics-based caching strategies outperform significantly LRU, with the higher relative and absolute performance gain noticed for small and large cache sizes, respectively. SG-LRU II.A presents



(a) Local hit rates.



(b) Global hit rates.



(c) Total hit rates.

Figure 3.5 Cache hit rates for varying number of user requests N_r.

the best performance across the whole relevant range of C. Similar observations are obtained from Fig. 3.7c, where the catalog size is varied, maintaining the same catalog size / cache size ratios as in the previous test.

As depicted in Fig. 3.8a, LRU II, SG-LRU, and SG-C I approach their baseline performance that is achieved for $N_r = 1,000,000$ already when $N_r = 100,000$, while WLFU-NE and SG-C II require 500,000 requests (in fact, for 100,000 requests the performance of SG-C II is LRU-like) and LRU I needs 1,000,000 requests.

Finally, in Fig. 3.8b is shown that the performance of all statistics-based strategies improves with the window size up to $W = 100,000$ and then remains constant or increases slightly for $W = 1,000,000$, with the exception of SG-C I whose performance degrades significantly. We note that for small window size, all caching strategies present an LRU-like behavior. From $W = 1,000$ and onwards, their performance starts to improve, with WLFU-NE presenting a slower rate of improvement in comparison to the other caching schemes. The best performance is achieved for large window size by SG-LRU I, SG-C I (with the exception of the scenario where $W = 1,000,000$), and WLFU-NE I.

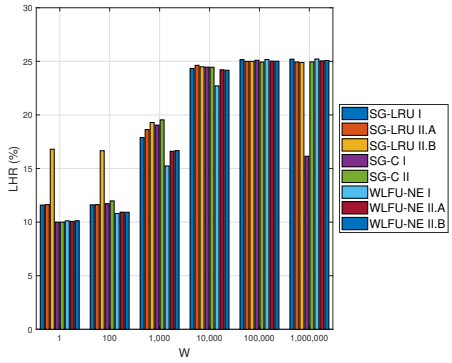
In summary, we note that the statistics-based strategies create a significant number of JT opportunities for moderate and large values of β, C, and W, especially for caches that deal with a large number of user requests.

3.6.3 Sum Spectral Efficiency

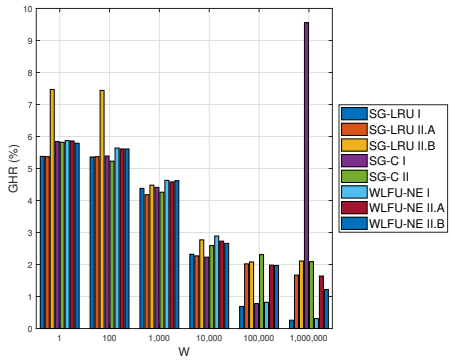
In this section we study the performance of the proposed resource allocation techniques. We consider a setup where $M = 2$, $K = 1$, and $N = 4$. We assume the use of SG-LRU I under a scenario where $\beta = 0.99$, $C = 100$, $F = 10,000$, $N_r = 1,000,000$, and $W = 100,000$, and we compare the performance (in terms of the average sum-SE achieved after $N_{\text{sim}} = 1,000$ simulation runs) of CoMP-CP vs. the performance of a hybrid scheme where CoMP-CP is utilized only when cache-aided CoMP-JT cannot take place (that is, about 74% of the time for this caching scenario). The application of C-ICEPA-ZF is considered in both cases. Fig. 3.9 depicts the simulation results. We note that the hybrid method slightly outperforms CoMP-CP for relaxed IPT and performs slightly worse for more tight IPT values.

3.7 Summary and Conclusions

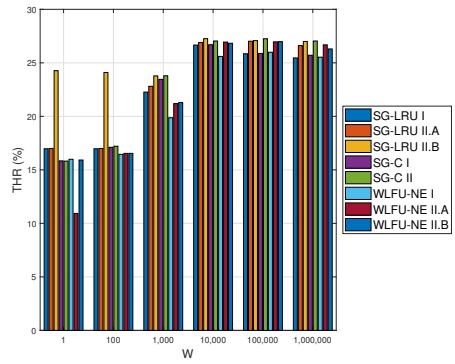
In this work, we presented a coordinated caching strategy and a number of statistics-based caching schemes. The latter achieve higher LHR and THR than the "de facto" LRU caching method and create more JT opportunities, especially for large β, C, and W, whereas LRU achieves better GHR. The use of cooperative caching variants that utilize global statistics improves the GHR. Finally, we saw that a hybrid CoMP-CP / cache-aided CoMP-JT strategy performs slightly better than or close to the CoMP-CP approach, assuming the application of C-ICEPA-ZF.



(a) Local hit rates.

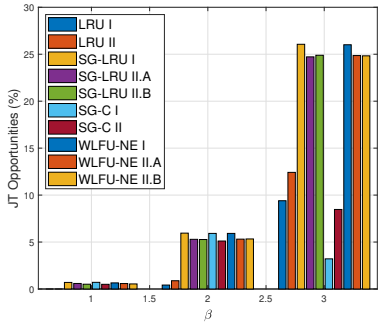


(b) Global hit rates.

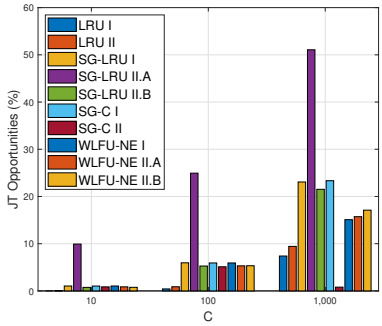


(c) Total hit rates.

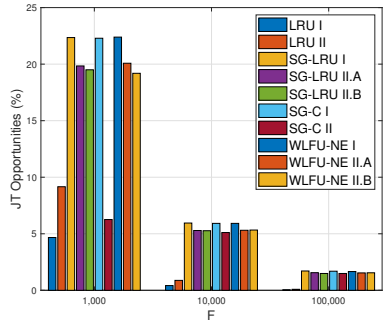
Figure 3.6 Cache hit rates for varying window size W.



(a) Varying Zipf shape parameter β.

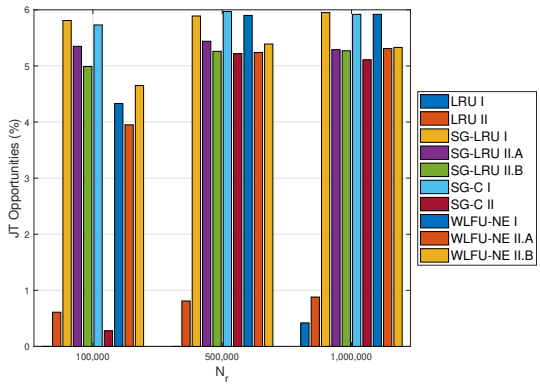


(b) Varying cache size C.

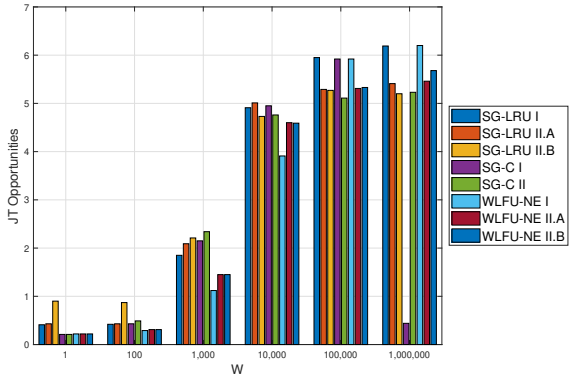


(c) Varying catalog size F.

Figure 3.7 Percentage of cache-aided JT opportunities for varying β, C, and F.



(a) Varying number of requests N_r.



(b) Varying window size W.

Figure 3.8 Percentage of cache-aided JT opportunities for varying N_r and W.

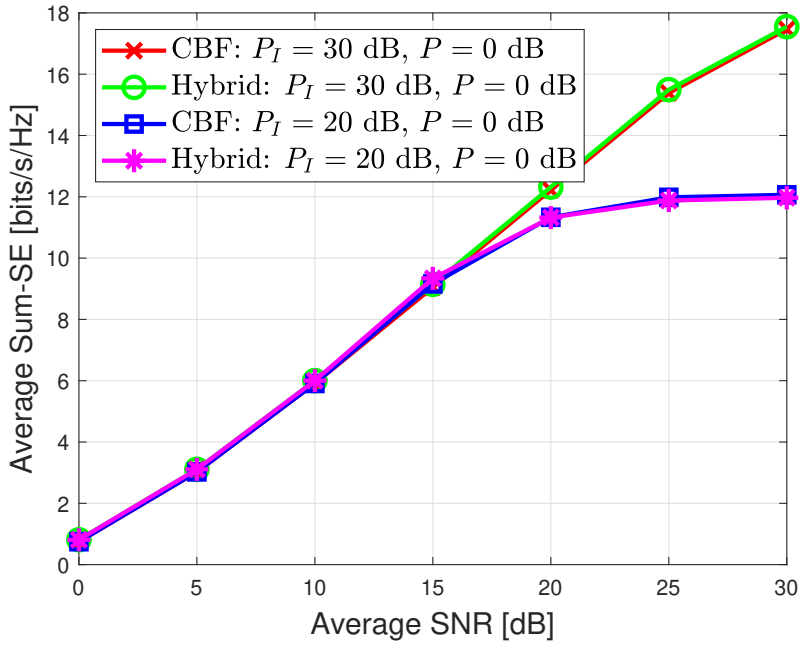


Figure 3.9 Average sum-SE vs. average SNR for CoMP-CP vs. hybrid CoMP-CP / CoMP-JT assuming the application of SG-LRU I and C-ICEPA-ZF.

Chapter 4

Spectrum Sharing III: Coordinated Hybrid Precoding

4.1 Introduction

The performance of CoMP transmission techniques improves with the number of base station (BS) antennas [128]. When a digital transceiver architecture is adopted, though, wherein each antenna element (AE) is fed by a radio frequency (RF) chain, adding more antennas increases the cost and power consumption [100]. Hybrid analog-digital transceivers use a mixture of active and passive AEs (AAE / PAE) to improve the array gain for a given number of RF units [100, 101, 128, 129] (see Fig. 4.1).

However, the use of such transceivers, which utilize a combination of digital (baseband) precoding and analog beamforming that is called hybrid precoding [129], has been studied almost exclusively in the very large antenna arrays (or massive MIMO) regime [27, 128].

In addition, the majority of the relevant works considers setups that make use of phased antenna arrays, although alternative architectures that utilize load-controlled parasitic antenna arrays (LC-PAA) further reduce the hardware cost and present compact designs, thus facilitating the installation at small-cell BSs and remote radio units [100, 115, 122, 128, 130–133]. This is because of the challenges in applying arbitrary hybrid precoding at the latter type of transceivers, e.g., due to the discrete values of the adjustable loads and the non-linear loads-currents relation [100, 115, 122, 128, 130–133].

In [134] is presented a precoding method for single-RF LC-PAAs, which is based on the mapping of the precoded signals onto the antenna currents and the calculation of the loading values (impedances) that generate these currents. The main issue with this approach is that it often requires the use of complex adjustable

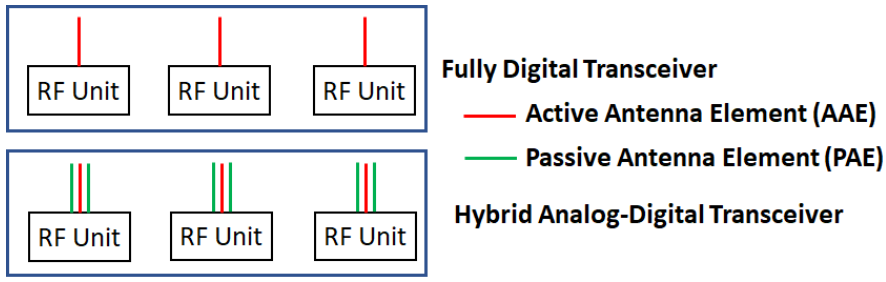


Figure 4.1 The concepts of fully digital and hybrid analog-digital transceivers.

loads whose real part ranges from negative to positive values [135]. A circuit implementation of such loads can be found in [136].

In [137, 138] the authors present hybrid precoding methods for single-RF LC-PAAs that are based on the approximation of the precoded signal by another signal that can be generated by simple analog loads, such that the approximation error is minimum (e.g., in the mean square sense [137]), to overcome the aforementioned issue.

In this chapter, we extend the work of [134] for coordinated linear precoding setups with either single-RF or multiple-RF LC-PAAs. In addition, we describe an alternative beam selection and precoding (BSP) method to overcome the load computation issues [115, 122, 130–133]. Finally, we present the application of coordinated hybrid symbol-level zero-forcing (ZF) precoding, which provides performance gains in the noise-limited low SNR regime, at LC-PAA-equipped setups [115, 130, 132].

4.2 Load-Controlled Parasitic Antenna Arrays

A load-controlled parasitic antenna array (LC-PAA) is comprised by a limited number of AAEs (i.e., AEs that are fed by an RF unit) surrounded by a large number of PAEs [100, 115, 122, 128, 130–133]. The latter AEs are deliberately placed in close vicinity to the AAEs and are terminated to tunable analog loads, such as varactors [100, 115, 122, 128, 130–133]. As a consequence of the strong mutual coupling among the antennas, the feeding voltages induce currents on the ports of the PAEs, thus enabling them to radiate [100, 115, 122, 128, 130–133]. The far-field radiation pattern of the antenna array is the superposition of the individual antenna responses [100, 115, 122, 128, 130–133]. By adjusting the baseband weights and the currents on the parasitic antennas (e.g., via an inexpensive digital control circuit), we can perform hybrid precoding [100, 115, 122, 128, 130–133]. Note that in contrast to phased antenna arrays, LC-PAAs do not utilize feeding networks and phase shifting modules, neither require a sufficiently large inter-element dis-

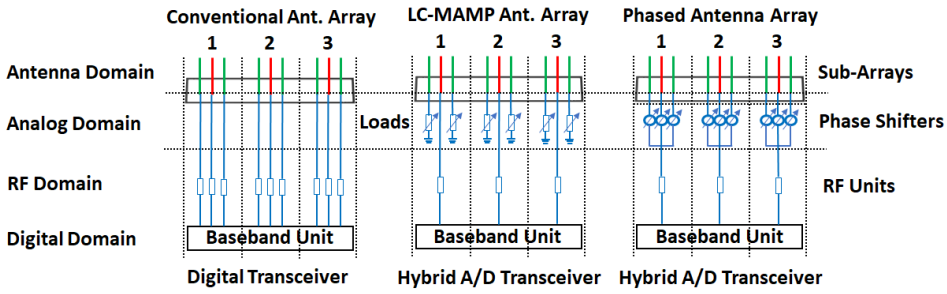


Figure 4.2 Structure of a fully digital transceiver and of hybrid analog-digital transceivers based on a LC-MAMP or a phased antenna array.

tance to avoid the occurrence of electromagnetic coupling among the AEs—in contrast, they exploit the mutual coupling to enable adapting beamforming / precoding with smaller number of RF units than antennas[1] [100, 115, 122, 128, 130–133]. Therefore, LC-PAAs represent more cost effective and compact architectures than phased antenna arrays [100, 115, 122, 128, 130–133]. A multi-RF LC-PAA is called a load-controlled multiple-active multiple-passive (LC-MAMP) antenna array [100, 115, 122, 128, 130–133]. The special case of a single-RF LC-PAA is referred to as a load-controlled single-active multiple-passive (SAMP) antenna array [100, 115, 122, 128, 130–133]. Fig. 4.2 compares the structure of a fully digital transceiver, a hybrid analog-digital transceiver that utilizes a LC-MAMP, and another hybrid analog-digital transceiver that utilizes a phased antenna array.

4.3 Coordinated Precoding with LC-MAMP Arrays

The signal model for a multi-user MIMO setup where a BS that is equipped with a N-antennas LC-MAMP array serves K single-antenna mobile stations (MS) is given by [115, 122, 128, 130–133]:

$$\mathbf{y} = \mathbf{H}\mathbf{i} + \mathbf{n}, \tag{4.1}$$

where $\mathbf{y}$ is the vector of open-circuit voltages at the receive antennas, $\mathbf{i}$ is the vector of currents that run on the transmit antennas, $\mathbf{H}$ is the channel matrix that relates these quantities with each other, and $\mathbf{n}$ is the receive additive noise vector.

However, we know that a transmission over such a MIMO broadcast channel (BC) is represented by [8]:

$$\mathbf{y} = \mathbf{H}\mathbf{W}\mathbf{s} + \mathbf{n}, \tag{4.2}$$

where $\mathbf{W}$ is the precoding matrix and $\mathbf{s}$ is the transmitted symbols vector. Hence,

[1] The inter-element distance between AAEs should be larger than $\lambda/2$, as usual, but the inter-element distance between PAEs is typically between $\lambda/10$ and $\lambda/20$ [100].

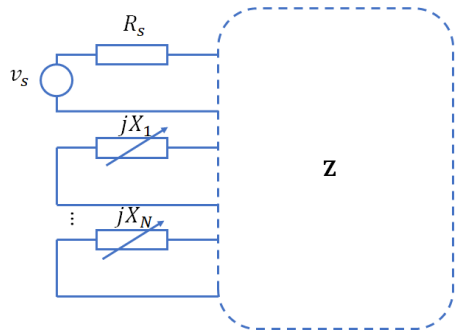


Figure 4.3 Equivalent circuit diagram of a LC-SAMP antenna array.

we can perform arbitrary channel-dependent precoding with LC-MAMPs by mapping the precoded symbols onto the antenna currents [115, 122, 128, 130–133]. That is [115, 122, 128, 130–133],

$$\mathbf{i} = \mathbf{W}\mathbf{s}. \tag{4.3}$$

Fig. 4.3 illustrates the equivalent circuit diagram of a load-controlled single-active multiple-passive (LC-SAMP) antenna array [115, 122, 128, 130–133]. We see that [115, 122, 128, 130–133]:

$$\mathbf{i} = (\mathbf{Z} + \mathbf{X})^{-1}\mathbf{v}, \tag{4.4}$$

where $\mathbf{Z}$ is the mutual impedance matrix. The diagonal elements of $\mathbf{Z}$, Z_{ii}, represent the self-impedance of the i-th AE, whereas its off-diagonal elements, Z_{ij}, denote the mutual impedance between the i-th AE and the j-th AE $(i,\ j = 1,\ldots,N)$. X is the diagonal load matrix that holds on its main diagonal the source resistances and the impedances of the analog loads that are connected to the PAEs. Finally, $\mathbf{v}$ is the voltage vector whose non-zero elements are the source voltages.

We can rearrange Eq. (4.4) as follows [128, 138]:

$$(\mathbf{Z} + \mathbf{X})\,\mathbf{i} = \mathbf{v}. \tag{4.5}$$

That is [128, 138],

$$\begin{bmatrix} Z_{11} + X_1 & \cdots & Z_{1N} \\ \vdots & \ddots & \vdots \\ Z_{N1} & \cdots & Z_{NN} + X_N \end{bmatrix} \begin{bmatrix} i_1 \\ \vdots \\ i_N \end{bmatrix} = \begin{bmatrix} v_1 \\ \vdots \\ v_N \end{bmatrix}. \tag{4.6}$$

Since for given $\mathbf{W}$ and $\mathbf{s}$ the vector of the currents $\mathbf{i}$ is obtained from Eq. (4.3), then assuming a LC-MAMP antenna array with N_a AAEs and N_p PAEs (i.e., $N = N_a + N_p$ AEs in total) where $X_n = 50$ for $n = 1,\ldots,N_a$ and $v_n = 0$ for $n = N_a +$

$1,\ldots,N$, we can solve Eq. (4.6) for the source voltages v_n and the impedances X_n [128]:

$$v_n = \sum_{j=1}^{N} Z_{nj} i_j + 50 i_n, \quad n = 1,\ldots,N_a. \tag{4.7a}$$

$$X_n = \frac{-\sum_{j=1}^{N} Z_{nj} i_j}{i_n}, \quad n = N_a + 1,\ldots,N. \tag{4.7b}$$

This approach allows us to perform hybrid precoding with LC-MAMP antenna arrays [115, 122, 128, 130–133]. The extension to coordinated hybrid precoding is trivial, e.g., for coordinated ZF precoding, the (non-normalized) precoder of the m-th BS is $\mathbf{F}_m^{(\mathrm{ZF})} = (\mathbf{H}_m)^{\#}$, as shown in Chapter 2 [115, 122, 128, 130–133].

4.4 Beam Selection and Precoding

In this section, we present a "divide-and-conquer" approach that enables us to avoid the use of complex loads as well as dynamic load computation all together for performing channel-dependent precoding [115, 122, 130–133]. This method, which is called beam selection and precoding (BSP), exploits the transmit beamforming (BF) capabilities of LC-PAAs and is described as follows [115, 122, 130–133]:

1. First, transmit BF is applied in the analog domain (loads) using any valid method.
2. Then, precoding is applied in the digital domain (baseband).
3. Finally, the precoded signals are transmitted over the employed beams.

This technique takes advantage of the fact that the required array manifold does not depend on the input signal [115, 122, 130–133]. The loading values in this case play simply the role of BF weights, i.e., they determine the amplitude and phase of the currents on the parasitic antennas, so that the required radiation pattern is shaped [115, 122, 130–133]. This reconfigurability of LC-PAAs in the analog domain enables us to decouple the problem into an analog BF part and a digital precoding part [115, 122, 130–133].

Let us describe BSP in more detail. Consider a setup with M BSs and K active single-antenna MS in each cell. Each BS is equipped with a LC-MAMP with $N = N_a + N_p$ AEs and at each timeslot can generate one out of b distinct precomputed beams, where each one corresponds to a different set of loading values [115, 122, 130–133]. Hence, there are b^M possible beam combinations (M-tuples of beams) [115, 122, 130–133]. The system operation is divided in three phases [115, 122, 130–133]:

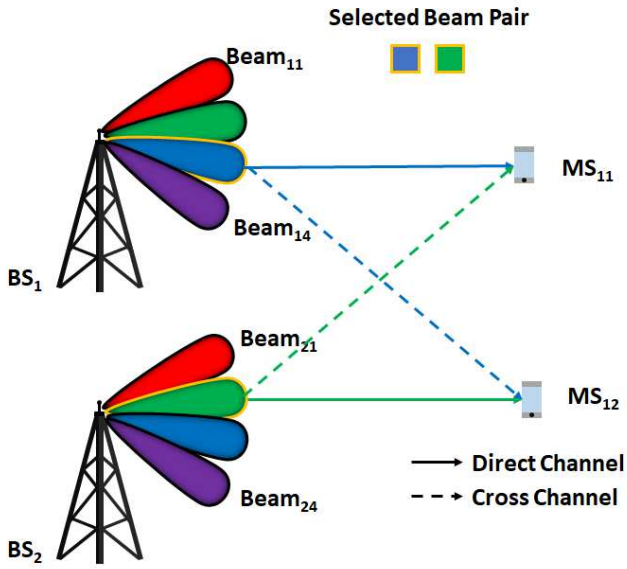


Figure 4.4 Beam selection.

1. **Learning phase:** Each BS cycles through its beams and sends a pilot for each one of them. The corresponding MS reports back the channel state information (CSI) or the signal-to-interference-plus-noise-ratio (SINR), as an alternative with low feedback overhead.

2. **Beam selection phase:** Then, the best beam pair, in terms of the achieved sum-rate (SR), is selected.

3. **Precoding and transmission phase:** Finally, the data signals are precoded in the digital domain and the precoded signals are transmitted over the selected beams. Note that if beam selection was based on SINR feedback, then CSI feedback takes place for the selected "beam channel", in order for precoding to be applied.

Fig. 4.4 shows the beam selection concept for a system setup with $M = 2$ cells, $K = 1$ user per cell, and $b = 4$ beams per BS [115, 122, 130–133].

The number of predetermined beams represents in a sense the spatial resolution of the system. It defines also the CSI or SINR acquisition overhead [115, 122, 130–133].

4.5 Coordinated Symbol-Level Precoding

The conventional codeword-level precoding schemes aim at mitigating or even eliminating the intra-cell and inter-cell co-channel interference (CCI) [115, 130, 132]. At symbol-level, though, CCI may be constructive, meaning that it may enhance the receive signal-to-noise-ratio (SNR) [115, 130, 132]. Symbol-level ZF precoding—also known as constructive interference ZF (CIZF) precoding—exploits this fact: it cancels the destructive symbol-level CCI and leaves unaffected the constructive one, in order to improve the performance in the noise-limited low SNR regime [115, 130, 132, 139].

Let us define the channel cross-correlation matrix $\mathbf{R} \in \mathbb{C}^{K_T \times K_T}$ [115, 130, 132]:

$$\mathbf{R} = \mathbf{H}\mathbf{H}^{\dagger}, \tag{4.8}$$

where $\mathbf{H} \in \mathbb{C}^{K_T \times N_T}$ holds the channels from all BSs to all users.

Next, let us index the users as $1,\ldots,K,K+1,\ldots,2K,\ldots,(M-1)K,\ldots,MK$. Then, the symbol-to-symbol interference from s_k to s_m is expressed as [115, 130, 132]:

$$I_{km}^{sym} = s_k \rho_{km}^{sym}, \quad m \neq k, \tag{4.9}$$

while the cumulative interference on s_m from all symbols is [115, 130, 132]:

$$I_{km,T}^{sym} = s_k \sum_{\substack{m=1 \\ m \neq k}}^{K_T} \rho_{km}^{sym}, \tag{4.10}$$

where

$$\rho_{km}^{sym} = \frac{(\mathbf{h}_k)^{\dagger}\mathbf{h}_m}{\mathbf{h}_k \mathbf{h}_m} \tag{4.11}$$

is the (k,m) entry of $\mathbf{R}$ and $\mathbf{h}_i$ is the direct channel that supports the transmission of s_i $(i = 1,\ldots,K_T)$.

We define the composite symbol-level ZF precoding matrix as [115, 130, 132]:

$$\mathbf{W}^{(\mathrm{CIZF})} = \mathbf{W}^{(\mathrm{ZF})}\mathbf{T} = \mathbf{H}^{\dagger}\mathbf{R}^{-1}\mathbf{T}, \tag{4.12}$$

where under this notation $\mathbf{W}^{(\mathrm{ZF})} = \mathbf{H}^{\#}$.

The received signal at the k-th user is [115, 130, 132]:

$$y_k = \tau_{kk}\sqrt{p}_k s_k + \sum_{m \neq k} \tau_{km}\sqrt{p}_m s_m + n_k, \quad k,\ m = 1,\ldots,K_T, \tag{4.13}$$

where $\tau_{km}\sqrt{p}_m s_m$ is the CI from the m-th user to the k-th user. Thus, the SINR of

the k-th user is given by [115, 130, 132]:

$$\gamma_k = \sum_{m=1}^{K_T} |\tau_{km}|^2 p_m, \quad k = 1, \ldots, K_T. \tag{4.14}$$

The calculation of $\mathbf{T} \in \mathbb{C}^{K_T \times K_T}$ depends on the applied modulation scheme [115, 130, 132]. Assuming binary phase shift keying (BPSK)[2], $\mathbf{T}$ is computed as follows [115, 130, 132]: First, we compute the matrix $\mathbf{G} \in \mathbb{C}^{K_T \times K_T}$ as

$$\mathbf{G} = \text{diag}(\mathbf{s}) \text{Re}(\mathbf{R}) \text{diag}(\mathbf{s}). \tag{4.15}$$

Then, $\tau_{kk} = \rho_{kk}^{sym}$ and $\tau_{km} = 0$ if $g_{km} < 0$ or $\tau_{km} = \rho_{km}^{sym}$ otherwise.

Since CIZF is tightly coupled with the applied modulation scheme, we cannot use the Shannon "$\log_2$" capacity formula; we should take into consideration in the capacity computation the actual signal constellation instead [115, 130, 132]. Under this context, the data rate of the k-th user is given by [115, 130, 132]:

$$R_k = (1 - \text{BLER}) m, \tag{4.16}$$

where $m = 1$ symbol for BPSK. BLER is the block error rate and is given by [115, 130, 132]:

$$\text{BLER} = 1 - (1 - P_e)^{N_f}, \tag{4.17}$$

where P_e is the symbol error rate (SER)—and bit error rate (BER) for the BPSK case—and N_f is the frame size.

4.6 Performance Evaluation

In this section, we assess the performance of coordinated codeword-level and symbol-level hybrid precoding with LC-PAAs in underlay spectrum sharing setups via numerical simulations. In Fig. 4.5 is compared the performance of C-IUPA-ZF vs. the performance of C-ICPA-ZF for $P_I = \{30 \text{ dB}, 15 \text{ dB}\}$ and $P = \{0 \text{ dB}, 10 \text{ dB}\}$ assuming a setup with $M = 2$ cells and $K = 1$ single-antenna MS in each cell. Each BS is equipped with a LC-SAMP with $N_p = 10$ PAEs. The precoding matrix has been computed according to the framework described in Section 4.3. We note that the slightly increased array gain of the LC-SAMP in comparison to a corresponding single-RF directional antenna array improves slightly the performance.

Fig. 4.6 illustrates the performance of C-ICPA-RZF, C-ICPA-ZF, and C-ICPA-MRT with CSI-based and SINR-based beams selection in the same setup as before, assuming $P_I = 15$ dB, $P = 0$ dB, and $b = 4$. We notice that beam selection results

[2]CIZF has been designed to improve the performance in the low SNR regime, where it does not make sense to use high order modulation schemes.

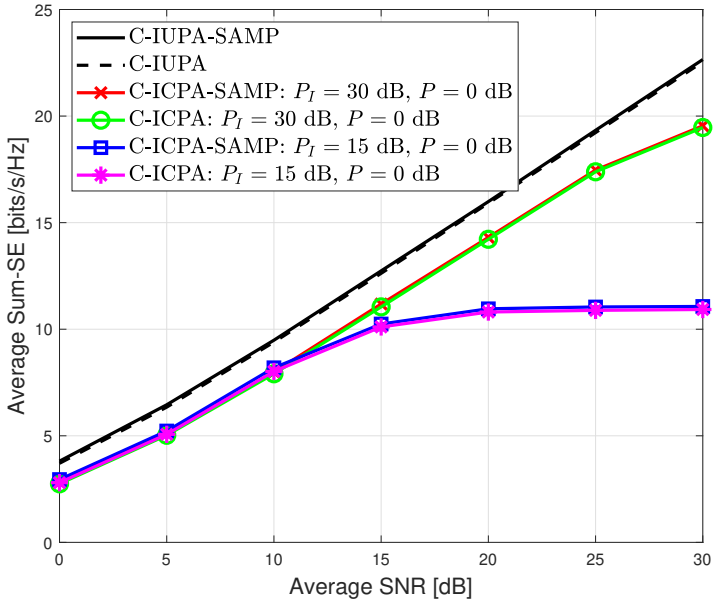


Figure 4.5 C-IUPA-ZF vs. C-ICPA-ZF for various P_I and P values in a setup with $M = 2$, $K = 1$, $N = 5$, $N_a = 1$, and $N_p = 4$.

in capacity reduction in comparison to the arbitrary channel-dependent precoding framework. Furthermore, we see that RZF with CSI-based beam pair selection outperforms its ZF and MRT counterparts, as expected. Finally, we note that SINR-based beam pair selection leads to a significant performance loss.

Fig. 4.7 shows the performance of C-ICPA-CIZF and C-ICPA-ZF for the same setup, assuming $P_I = 15$ dB and $P = 0$ dB and considering the use of BPSK modulation and a frame size of $N_f = 100$ symbols. We note that CIZF outperforms its codeword-level counterpart in the noise-limited low SNR regime, as expected.

4.7 Summary and Conclusions

In this chapter, we studied coordinated precoding with LC-PAAs in underlay spectrum sharing setups. We presented a framework that enables us to calculate the loads and digital weights required to apply arbitrary channel-dependent coordinated hybrid precoding at BSs equipped with single-RF or multi-RF LC-PAAs. The simulations indicated that these antenna arrays have the ability to slightly improve the performance in comparison to equivalent fully-digital systems. However, the implementation of arbitrary channel-dependent precoding with such hy-

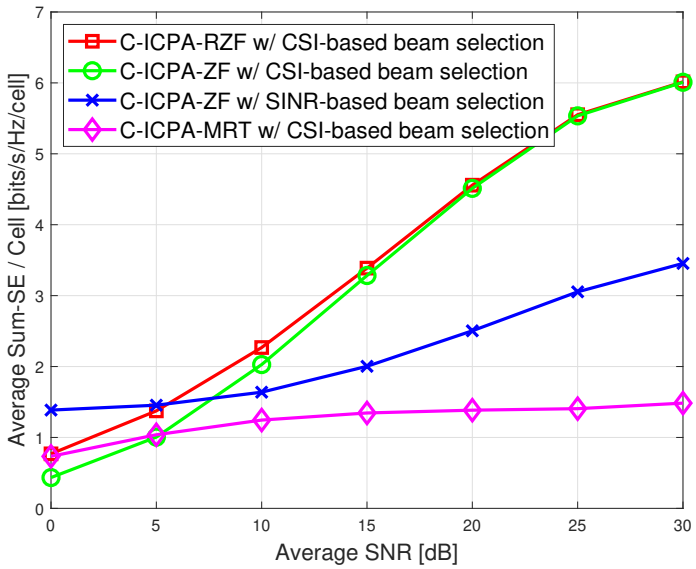


Figure 4.6 Beam-selection variants for a setup with $P_I = 30$ dB, $P = 0$ dB, $M = 2$, $K = 1$, $N = 5$, $N_a = 1$, and $N_p = 4$.

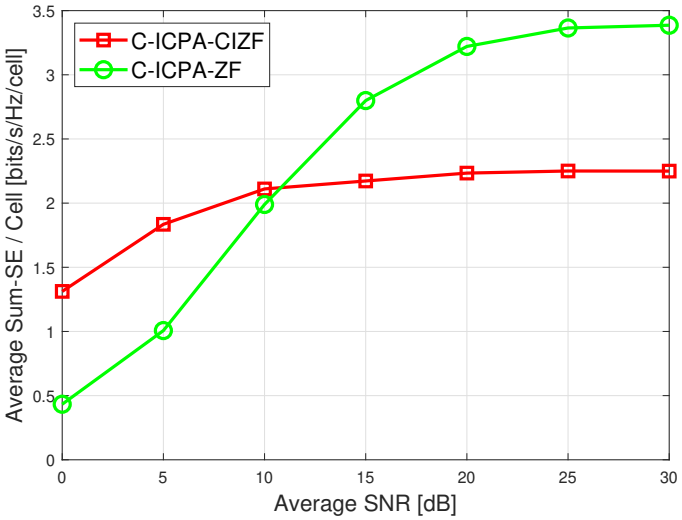


Figure 4.7 C-ICPA-CIZF vs. C-ICPA-ZF for $P_I = 15$ dB, $P = 0$ dB.

brid analog-digital transceivers is challenging, hardware-wise. Thus, we presented also a beam selection and precoding (BSP) method, where the loads are utilized only for analog beamforming and precoding is performed in the digital domain (baseband), to address this issue. Nevertheless, this approach results in substantial performance loss if a small number of precomputed beams is utilized—and the number of beams is limited in practice by the learning overhead. Finally, we described the application of coordinated symbol-level ZF precoding at LC-PAAs. The simulation results demonstrated that this scheme improves the performance in the low SNR regime.

Chapter 5

Spectrum Sharing IV: Hybrid Precoding for Massive MIMO

5.1 Introduction

Massive MIMO (mMIMO) has been recognized as an integral component of next-generation cellular networks, due to its high capacity gain that is attributed to the excess number of spatial degrees-of-freedom (DoF) [56]. Moreover, it is considered a key enabler of millimeter-wave (mmWave) wireless access, since its high array gain compensates for the severe distance-dependent path loss in such high frequencies [1]. At the same time, the small wavelengths enable the packing of hundreds of antennas in compact antenna arrays [1]. However, the cost and power consumption of fully digital transceivers is high (especially in mmWave implementations), since in these architectures each antenna element (AE) is fed by its own radio frequency (RF) chain [100].

Hybrid analog-digital designs have been proposed as a workaround to this problem. These transceiver architectures exploit the fact that the multiplexing gain is determined by the number of RF units, whereas the array gain is determined by the total number of AEs [101]. Typically, hybrid designs for mMIMO consist of a limited number of RF modules that are connected to a large number of antennas via phase shifters and perform precoding on the digital domain and beamforming on the analog domain (see Fig. 5.1). These structures provide a good balance between performance vs. cost and power consumption. The performance loss is attributed to the limitations of analog processing, since the phase shifters impose a constant amplitude constraint [1].

Hybrid processing techniques have been proposed for both sub-6 GHz and mmWave channels. Spatially sparse precoding via orthogonal matching pursuit (SSPOMP) [102], which explores the sparsity (i.e., poor scattering) of the mmWave

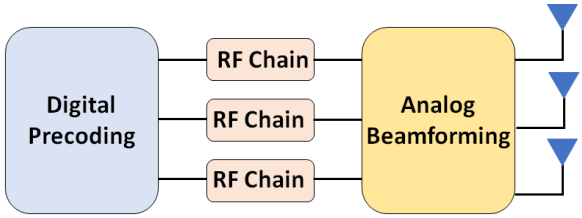


Figure 5.1 Hybrid analog-digital transceiver [1].

channel, constitutes the state-of-the-art in mmWave hybrid processing. Nevertheless, this technique cannot be applied for rich scattering (i.e., Rayleigh) environments.

In this work, we propose an efficient algorithm for hybrid processing that can be applied in both rich (Rayleigh) and poor (mmWave) scattering environments. The proposed method alternates between the optimization of the baseband and the analog precoder/combiner. For the latter, a *convolution smoothing technique* [140] that avoids local minima followed by a *stochastic approximation scheme* is employed for the estimation of the phases. The advantage of the proposed method is that it demonstrates very good performance with low computational complexity, especially in rich scattering environments.

5.2 Signal Model

We consider a mMIMO link that is established between a hybrid transmitter equipped with N_t antennas and M_t RF chains and a hybrid receiver with N_r antennas and M_r RF chains. The link supports N_s data streams. The architecture of both systems is a fully connected one with fewer RF chains than antennas, where $N_s \leq M_t \leq N_t$ and $N_s \leq M_r \leq N_r$. The transmitter applies a $M_t \times N_s$ baseband precoder $\mathbf{F}_B$ (enabling both amplitude and phase modifications) and a $N_t \times M_t$ analog precoder $\mathbf{F}_R$ (enabling phase changes only). Therefore, each (i,j)-th element of $\mathbf{F}_R$ satisfies $\left|(\mathbf{F}_R)_{i,j}\right| = 1/\sqrt{N_t}$. Finally, to meet the total transmit power constraint $\mathbf{F}_B$ is normalized to satisfy $\|\mathbf{F}_R\mathbf{F}_B\|_F^2 = N_s$.

The received signal $\mathbf{y} \in \mathbb{C}^{N_r \times 1}$ before combining is given by [141]

$$\mathbf{y} = \mathbf{H}\mathbf{F}_R\mathbf{F}_B\mathbf{s} + \mathbf{n}, \tag{5.1}$$

where $\mathbf{H} \in \mathbb{C}^{N_r \times N_t}$ is the normalized channel matrix with $\mathbb{E}[\|\mathbf{H}\|_F^2] = N_t N_r$, $\mathbf{s} \in \mathbb{C}^{N_s \times 1}$ is the transmitted signal, and $\mathbf{n} \sim \mathcal{CN}(\mathbf{0}, \sigma_n^2 \mathbf{I}_{N_r})$ is the i.i.d. noise vector. The average transmit power is P.

Assuming CSIR and CSIT, the received signal after hybrid combining is [141]:

$$\widetilde{\mathbf{y}} = \mathbf{W}_B^{\dagger}\mathbf{W}_R^{\dagger}\mathbf{H}\mathbf{F}_R\mathbf{F}_B\mathbf{s} + \mathbf{z}, \tag{5.2}$$

where $\mathbf{z} = \mathbf{W}_B^{\dagger}\mathbf{W}_R^{\dagger}\mathbf{n}$, while $\mathbf{W}_R$ denotes the $N_r \times M_r$ analog combining matrix and $\mathbf{W}_B$ is the $M_r \times Ns$ baseband combining matrix. A fully connected phase shifter design is considered for the combiner as well, hence, $\left|(\mathbf{W}_R)_{i,j}\right| = 1/\sqrt{N_r}$. If Gaussian signaling is employed and equal power allocation (EPA) is considered, then the achieved instantaneous spectral efficiency (SE) is expressed via [141]:

$$R(\mathbf{F}_R, \mathbf{F}_B, \mathbf{W}_R, \mathbf{W}_B) = \log_2(|\mathbf{I}_{N_s} + \frac{P}{N_s}\mathbf{R}_{\mathbf{z}}^{-1}\widetilde{\mathbf{H}}\widetilde{\mathbf{H}}^{\dagger}|), \tag{5.3}$$

where $\mathbf{R}_n = \sigma_n^2\mathbf{W}_B^{\dagger}\mathbf{W}_R^{\dagger}\mathbf{W}_R\mathbf{W}_B$ is the covariance matrix of the noise and $\widetilde{\mathbf{H}} = \mathbf{W}_B^{\dagger}\mathbf{W}_R^{\dagger}\mathbf{H}\mathbf{F}_R\mathbf{F}_B$.

For limited scattering (sparse) we consider the geometric clustered mmWave model [102] with N_c clusters and N_p paths per cluster:

$$\mathbf{H} = \sqrt{\frac{N_t N_r}{N_c N_p}} \sum_{m=1}^{N_c}\sum_{n=1}^{N_p} \beta_{mn}\mathbf{a}_r(\phi_{mn})\mathbf{a}_t^{\dagger}(\theta_{mn}), \tag{5.4}$$

where $\beta_{mn} \sim \mathcal{CN}(0,1)$ is the complex channel gain of the (m,n)-th path. $\mathbf{a}_r(\phi_{mn})$ denotes the receive array response vector at the azimuth angle of arrival (AoA) ϕ_{mn} and $\mathbf{a}_t(\theta_{mn})$ denotes the transmit array response vector at the azimuth angle of departure (AoD) θ_{mn}. The mean angles of each cluster (center) are uniformly distributed and the angles within each cluster are distributed according to the truncated Laplace distribution with angular spreads $\sigma_\phi, \sigma_\theta$, respectively.

Uniform linear arrays (ULA) with half-wavelength spacing of the N antenna elements (N_t for the transmitter and N_r for the receiver) are considered in our simulations, with the array response vector given by:

$$\mathbf{a} = [1, e^{-\mathrm{j}\pi\sin\theta}, \ldots, e^{-\mathrm{j}\pi\sin\theta(N-1)}]^T/\sqrt{N}. \tag{5.5}$$

5.3 Preliminaries

5.3.1 Hybrid Design

Assuming that $\text{rank}(\mathbf{H}) \geq N_s$, the optimal precoder $\mathbf{F}_*$ and combiner $\mathbf{W}_*$ of a fully digital system can be found by the singular value decomposition (SVD) of the channel matrix $\mathbf{H} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^{\dagger}$, where $\mathbf{U}$ and $\mathbf{V}$ are $N_r \times N_r$ and $N_t \times N_t$ unitary matrices, respectively, and $\mathbf{\Lambda}$ is a $N_r \times N_t$ diagonal matrix with singular values in decreasing order on its diagonal. The optimal unconstrained precoder and com-

biner, for equal power allocation, is given by $\{\mathbf{F}_*, \mathbf{W}_*\} = \{\mathbf{V}_1, \mathbf{U}_1\}$, where $\mathbf{V}_1$ and $\mathbf{U}_1$ are obtained from $\mathbf{V}$ and $\mathbf{U}$ by extracting their first Ns columns, respectively.

The joint optimization of the hybrid precoders $\{\mathbf{F}_R, \mathbf{F}_B\}$ and combiners $\{\mathbf{W}_R, \mathbf{W}_B\}$ (global minimum of the joint design) is a challenging task, due to the non-convex constraints of the analog precoder and combiner. The adopted approach is to first design hybrid precoders, which are sufficiently close to the optimal ones, i.e., $\mathbf{F}_* = \mathbf{V}_1$, by solving the following optimization task:

$$\begin{aligned} &\min_{\mathbf{F}_R, \mathbf{F}_B} \|\mathbf{F}_* - \mathbf{F}_R \mathbf{F}_B\|_F^2, \\ &\text{s.t. } \mathbf{F}_R \in \mathcal{F}_R, \ \|\mathbf{F}_R \mathbf{F}_B\|_F^2 = N_s, \end{aligned} \tag{5.6}$$

where $\mathcal{F}_R = \{\mathbf{F}_R \in \mathbb{C}^{N_t \times M_t} : \left|(\mathbf{F}_R)_{i,j}\right| = 1/\sqrt{N_t}\}$ is the set of matrices with constant-magnitude entries. The fact that the error of the approximation in Eq. (5.6) is non-zero makes $\mathbf{U}_1$ no further optimal. The linear minimum mean square error (MMSE) combiner that achieves the maximum spectral efficiency for linear and separate detection of each data stream is given by:

$$\mathbf{W}_* = \frac{\sqrt{P}}{N_s} \left(\frac{P}{N_s} \mathbf{H} \mathbf{F}_R \mathbf{F}_B \mathbf{F}_B^\dagger \mathbf{F}_R^\dagger \mathbf{H}^\dagger + \sigma_n^2 \mathbf{I}_{N_r} \right)^{-1} \mathbf{H} \mathbf{F}_R \mathbf{F}_B. \tag{5.7}$$

Hence, given the set of optimized precoders and calculating the $\mathbf{W}_*$ from Eq. (5.7), the hybrid combiner can be obtained in a similar manner as the solution to the following task:

$$\min_{\mathbf{W}_R, \mathbf{W}_B} \|\mathbf{W}_* - \mathbf{W}_R \mathbf{W}_B\|_F^2, \text{ s.t. } \mathbf{W}_R \in \mathcal{W}_R, \tag{5.8}$$

where $\mathcal{W}_R$ is the set of complex $N_r \times M_t$ matrices with constant-magnitude entries.

5.3.2 Gaussian Smoothing of Matrix Variable Functions

The random matrix $\mathbf{S} \in \mathbb{R}^{N \times M}$ follows a matrix variate normal distribution or MVND, denoted as $\mathbf{S} \sim \mathcal{MN}_{N \times M}(\mathbf{M}, \mathbf{\Sigma}, \mathbf{\Psi})$, where $\mathbf{M} \in \mathbb{R}^{N \times M}$ is its mean, and $\mathbf{\Sigma} \in \mathbb{R}^{N \times N}$, $\mathbf{\Psi} \in \mathbb{R}^{M \times M}$ are positive definite matrices, if $\text{vec}(\mathbf{S}) \sim \mathcal{N}_{NM}(\text{vec}(\mathbf{M}), \mathbf{\Psi} \otimes \mathbf{\Sigma})$ [142]. The p.d.f. of $\mathbf{S}$ is given by:

$$p(\mathbf{S}|\mathbf{M}, \mathbf{\Sigma}, \mathbf{\Psi}) = \frac{e^{-\frac{1}{2}\text{tr}\left(\mathbf{\Psi}^{-1}(\mathbf{S}-\mathbf{M})\mathbf{\Sigma}^{-1}(\mathbf{S}-\mathbf{M})^T\right)}}{\sqrt{(2\pi)^{NM} \det(\mathbf{\Sigma})^M \det(\mathbf{\Psi})^N}}. \tag{5.9}$$

Let $\mathbf{M} = \mathbf{O}_{N \times M}$ (zero matrix) and $\mathbf{\Sigma} = \beta^2 \mathbf{I}_N, \mathbf{\Psi} = \gamma^2 \mathbf{I}_M$. Moreover, considering

$\mu = \beta\gamma$, Eq. (5.9) is written as:

$$p(\mathbf{S},\mu) = \frac{e^{-\frac{1}{2\mu^2}\|\mathbf{S}\|_F^2}}{\mu^{NM}\sqrt{(2\pi)^{NM}}}. \tag{5.10}$$

The smoothed approximation to the original function f with weighting Gaussian p.d.f., $p(\mathbf{S},\mu)$, can be expressed via their convolution given by:

$$\begin{aligned} f_\mu(\mathbf{X}) = (p * f)(\mathbf{X}) &= \int_{\mathbb{R}^{N\times M}} p(\mathbf{S},\mu) f(\mathbf{X}-\mathbf{S}) d\mathbf{S}, \\ &= \int_{\mathbb{R}^{N\times M}} p(\mathbf{S}) f(\mathbf{X}-\mu\mathbf{S}) d\mathbf{S}, \end{aligned} \tag{5.11}$$

where $p(\mathbf{S}) = p(\mathbf{S},1)$ is the standard MVND and with the use of the change of variables. From Eq. (5.11), it is directly observed that $f_\mu(\mathbf{X}) = \mathbb{E}_{\mathbf{S}}\left[f(\mathbf{X}-\mu\mathbf{S})\right]$, which leads to:

$$\nabla_{\mathbf{X}} f_\mu(\mathbf{X}) = \mathbb{E}_{\mathbf{S}}\left[\nabla_{\mathbf{X}} f(\mathbf{X}-\mu\mathbf{S})\right], \tag{5.12}$$

where i.i.d. samples are obtained from the $\mathbb{R}^{N\times M}$ space with the p.d.f. $p(\mathbf{S})$. Therefore, the (one-sided) unbiased gradient estimator is expressed as $\nabla_{\mathbf{X}} f_\mu(\mathbf{X}) = \frac{1}{L}\sum_{\ell=1}^{L} \nabla_{\mathbf{X}} f(\mathbf{X}-\mu\mathbf{S}_{[\ell]})$. Using the change of variables $\mathbf{S} = -\mathbf{Y}$ in Eq. (5.12), summing and solving w.r.t. the gradient we obtain the two-sided estimate of the gradient, given by:

$$\nabla_{\mathbf{X}} f_\mu(\mathbf{X}) = \frac{1}{2L}\sum_{\ell=1}^{L}\left[\nabla_{\mathbf{X}} f(\mathbf{X}+\mu\mathbf{S}_{[\ell]}) + \nabla_{\mathbf{X}} f(\mathbf{X}-\mu\mathbf{S}_{[\ell]})\right]. \tag{5.13}$$

It should be noted that Eq. (5.13) suggests that L samples can be used for the gradient estimation, as in a mini-batch approach. However, in this work we only consider its stochastic flavor, i.e., $L = 1$ [140].

5.4 Hybrid Precoding: Stochastic Approximation with Gaussian Smoothing

In this section we introduce an *iterative scheme* for Hybrid Precoding via Stochastic Approximation with Gaussian Smoothing (HPSAGS), which alternates between the optimization of the digital and the analog precoder, as a solution of Eq. (5.6). The solution of (5.8) for the combiners is similar and hence omitted (see Remark). Therefore, we drop the transmitter-receiver indices from the dimensions and use N, M instead. The scheme is summarized in Algorithm 5.1 and Algorithm 5.2.

Algorithm 5.1 Hybrid Precoding via Stochastic Approximation with Gaussian Smoothing

procedure HPSAGS($\mathbf{F}_*, \boldsymbol{\Theta}_0, (\mu_k)_{k=0}^{K-1}, \eta, T_{\max}, \epsilon$)
  Set $k \leftarrow 0$
  while $k < K$ **do**
    Select $\mu \leftarrow \mu_k$
    Compute $\mathbf{F}_B^{(k)}$ in Eq. (5.14) with $\mathbf{F}_R^{(k)} = g(\boldsymbol{\Theta}_k)$
    $\boldsymbol{\Theta}_{k+1} \leftarrow \text{SGDM}(\mathbf{F}_*, \boldsymbol{\Theta}_k, \mathbf{F}_B^{(k)}, \mu, \eta, T_{\max}, \epsilon)$ using Alg. 5.2
    Set $k \leftarrow k+1$
  end while
  Compute $\mathbf{F}_R^{(K)} = g(\boldsymbol{\Theta}_K)$ and $\mathbf{F}_B^{(K)}$ from Eq. (5.14)
  $\mathbf{F}_B^{(K)} \leftarrow \sqrt{N_s}\mathbf{F}_B^{(K)} / \|\mathbf{F}_R^{(K)}\mathbf{F}_B^{(K)}\|_F$
  Output: $\mathbf{F}_R^{(K)}, \mathbf{F}_B^{(K)}$
end procedure

5.4.1 Baseband Precoder Update

Given an initial solution $\mathbf{F}_R^{(0)}$, the set of hybrid precoders at the k-th iteration is $\left\{\mathbf{F}_R^{(k)}, \mathbf{F}_B^{(k)}\right\}$. Provided we have computed $\mathbf{F}_R^{(k)}$ the baseband precoder update, $\mathbf{F}_B^{(k)}$, is given by the solution of $\min_{\mathbf{F}_B} \left\|\mathbf{F}_* - \mathbf{F}_R^{(k)}\mathbf{F}_B\right\|_F^2$, expressed in close form as:

$$\mathbf{F}_B^{(k)} = \left(\left(\mathbf{F}_R^{(k)}\right)^\dagger \mathbf{F}_R^{(k)}\right)^{-1} \left(\mathbf{F}_R^{(k)}\right)^\dagger \mathbf{F}_*. \tag{5.14}$$

5.4.2 Analog Precoder Update via Stochastic Approximation with Gaussian Smoothing

Next, follows the update of the analog precoder. To this end, we impose the constant-modulus structure on the matrix. Considering the non-linear mapping $g : \mathbb{R}^{N\times M} \to \mathbb{C}^{N\times M}$ with $g(\boldsymbol{\Theta}) = e^{j\boldsymbol{\Theta}}/\sqrt{N}$, the precoder matrix is expressed via the element-wise function $\mathbf{F}_R = g(\boldsymbol{\Theta}), \boldsymbol{\Theta} \in \mathbb{R}^{N\times M}$. Hence, we seek for $\boldsymbol{\Theta}_{k+1}$ minimizing $f : \mathbb{R}^{N\times M} \to \mathbb{R}$ with $f(\boldsymbol{\Theta}) = \|\mathbf{F}_* - g(\boldsymbol{\Theta})\mathbf{F}_B^{(k)}\|_F^2$, leading to the update $\mathbf{F}_R^{(k+1)} = f(\boldsymbol{\Theta}_{k+1})$. Note that f defines a multiextremal mapping with respect to $\boldsymbol{\Theta}$, due to the existence of the non-convex function g. Thus, standard approaches to find a minimizer of f do not apply here. Nevertheless, the function is smooth and

Algorithm 5.2 Stochastic Gradient Descent

procedure SGD($\mathbf{F}_*, \mathbf{\Theta}_k, \mathbf{F}_B^{(k)}, \mu, \eta, T_{\max}, \epsilon$)
  Set $t \leftarrow 0$, $\varepsilon_t \leftarrow \infty$ and $\mathbf{\Theta}_k^{(t)} \leftarrow \mathbf{\Theta}_k$.
  while $t < T_{\max}$ and $\varepsilon_t > \epsilon$ **do**
    Draw one sample from $p(\mathbf{S}, 1)$ in Eq. (5.10).
    Compute the gradients at $\mathbf{\Theta}_k^{(t)} + \mu\mathbf{S}, \mathbf{\Theta}_k^{(t)} - \mu\mathbf{S}$ using Eq. (5.15)
    Compute $\nabla_{\mathbf{\Theta}} f_\mu(\mathbf{\Theta}_k^{(t)})$ in (5.13) with $L = 1$
    Update gradient $\mathbf{\Theta}_k^{(t+1)} \leftarrow \mathbf{\Theta}_k^{(t)} - \eta \nabla_{\mathbf{\Theta}} f_\mu(\mathbf{\Theta}_k^{(t)})$
    Set $\varepsilon_t \leftarrow \|\mathbf{\Theta}_k^{(t+1)} - \mathbf{\Theta}_k^{(t)}\|_F / \|\mathbf{\Theta}_k^{(t)}\|_F$ and $t \leftarrow t + 1$
  end while
  Output: $\mathbf{\Theta}_{k+1}$.
end procedure

with gradient:

$$\nabla_{\mathbf{\Theta}} f(\mathbf{\Theta}) = -2\mathrm{Re}\{\mathrm{j} g(\mathbf{\Theta}) \odot ((\mathbf{F}_* - g(\mathbf{\Theta})\mathbf{F}_B^{(k)})^* \mathbf{F}_B^{(k)^T})\}, \tag{5.15}$$

where $\mathrm{Re}\{\cdot\}$ denotes the real part of the complex input, $\odot$ is Hadamard (element-wise) product and $(\cdot)^*$ the is the conjugate of the matrix.

The objective of *convolution function smoothing* [140] is to represent f as a superposition of a uniextremal function and other multiextremal ones, which add some noise to the former, and perform minimization of the smoothed uniextremal function by filtering out the noise, eventually leading towards its global minimum. This is performed by generating a sequence of minimization runs, while reducing the amount of smoothing at the end of the cycle. For the smoothing of f we have employed the framework in Section 5.3.2 in order to obtain the derivative, therefore, instead of minimizing f we attempt to solve the following stochastic optimization task at every k-th step:

$$\min_{\mathbf{\Theta}} \left\{ f_{\mu_k}(\mathbf{\Theta}) = \mathbb{E}_{\mathbf{S}} \left[f(\mathbf{\Theta} - \mu_k \mathbf{S}) \right] \right\}, \tag{5.16}$$

where $\mathbf{S}$ is sampled from the standard MVND in Eq. (5.10) and the sequence $(\mu_k)_{k \in \mathbb{N}}$ is strictly decreasing with $\lim_{n \to \infty} \mu_k = 0$. However, in practice, a small finite number K is sufficient for the approximation. Finally, at the k-th iteration, Stochastic Gradient Descent (SGD) is employed for the phases' update. The computational efficiency of the algorithm is expressed in terms of the worst case complexity, which is $\mathcal{O}(NM^2KT_{\max})$. The full code can be found online in [143].

Remark: *Note that for the design of the combiner the same algorithm can be used with minor modifications, i.e., by replacing* $\mathbf{F}_*$ *with* $\mathbf{W}_*$ *in Eq. (5.7) and by neglecting the normalization of the baseband matrix in row 9 of Algorithm 5.1.*

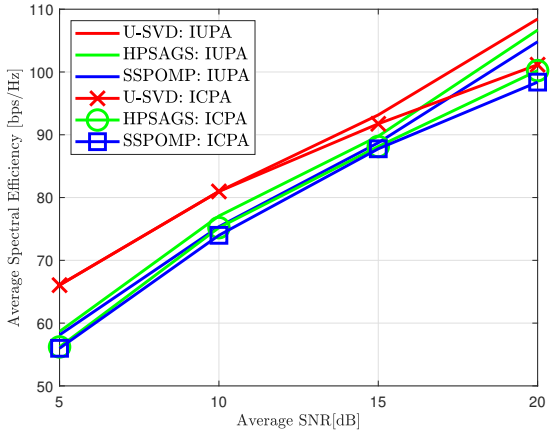


Figure 5.2 IUPA and ICPA for U-SVD, HPSAGS, and SSPOMP: $N_t = N_r = 64$, $M_t = M_r = 8$, $N_s = 8$, 64×64 primary link.

5.5 Performance Evaluation

In this section, we study the performance of interference-constrained power allocation (ICPA), assuming a mmWave mMIMO secondary link that coexists with another such primary link in an underlay spectrum sharing setup. The coexistence of a single-user (SU) MIMO link with another such link has been studied for fully digital transceivers in non-massive-MIMO setups operating at sub-6 GHz spectrum in [109], where the authors derived a SR maximization solution based on SVD precoding / combining and interference-constrained water-filling power allocation—a scheme that can be considered a special case of the coordinated ICPA method and its QoS-aware variant described in Chapter 2. In this chapter, we apply the ICPA solution for mmWave SU-mMIMO coexisting links.

Fig. 5.2 illustrates the performance of HPSAGS and SSPOMP for a hybrid mMIMO system with $N_t = N_r = 64$ antennas and $M_t = M_r = 8$ RF chains vs. the performance of unconstrained SVD (U-SVD) applied to an equivalent fully digital mMIMO system with $N_t = N_r = 64$ (which implies also 64 RF chains at the TX and the RX in this case) and $N_s = 8$ as well. We assume that $P_I = 20$ dB and $P = 0$ dB. The primary system is also a 64×64 mMIMO link. We note that HPSAGS slightly outperforms SSPOMP and approaches very closely U-SVD in the high SNR regime, although the hybrid system utilizes 8 times less RF units / node.

In Fig. 5.3 we repeat the previous test assuming $M_t = M_r = 12$ RF chains and $N_s = 12$ data streams. We see that in this case SSPOMP outperforms slightly HPSAGS. Also, we notice that the performance gap between the hybrid and digital processing solutions is slightly larger.

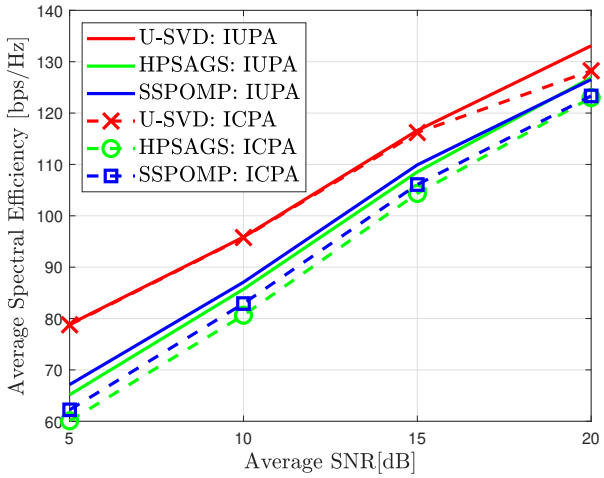


Figure 5.3 IUPA and ICPA for U-SVD, HPSAGS, and SSPOMP: $N_t = N_r = 64$, $M_t = M_r = 12$, $N_s = 12$, 64×64 primary link.

In Fig. 5.4 we repeat the previous test but now we vary the IPT as $P_I = \{20\text{dB}, 10\text{dB}\}$. We see that there is a very small performance loss due to the more hard IPC, which is more severe in the high SNR regime for SSPOMP than for HPSAGS.

In Fig. 5.5 we set back $M_t = M_r = 8$, $N_s = 8$, and $P_I = 20$ dB, but this time we consider both a 64×64 and a 128×128 primary link. We notice that there is a small performance degradation when the number of antennas at the primary transmitter and receiver increases, which is more severe in the high SNR regime for SSPOMP.

Finally, in Fig. 5.6 we repeat the previous test, considering this time $N_t = N_r = 128$ instead of $N_t = N_r = 64$. We observe that the performance improves for a larger number of antennas on the secondary system, as expected.

5.6 Summary and Conclusions

In this chapter, we studied hybrid precoding for mmWave massive MIMO links in underlay spectrum sharing setups. We presented a hybrid processing via stochastic approximation with Gaussian smoothing (HPSAGS) method that achieves high performance with low computational cost. The proposed technique approaches the performance of a fully digital mmWave mMIMO link and outperforms SSPOMP. The performance of HPSAGS becomes slightly worse when we transmit more data streams, the IPT is harder, or the secondary link has more antennas.

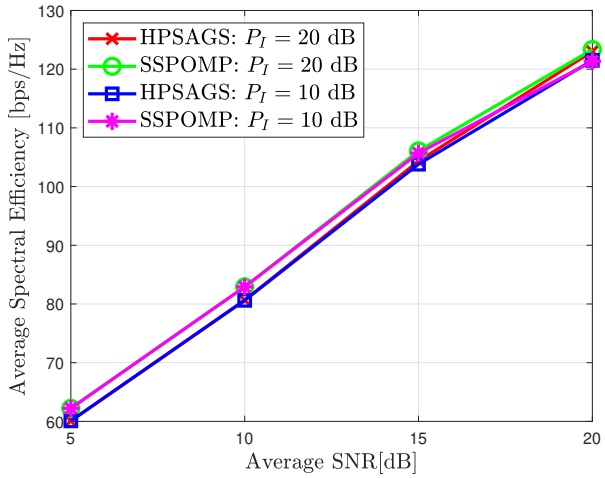


Figure 5.4 ICPA for U-SVD, HPSAGS, and SSPOMP with varying IPT.

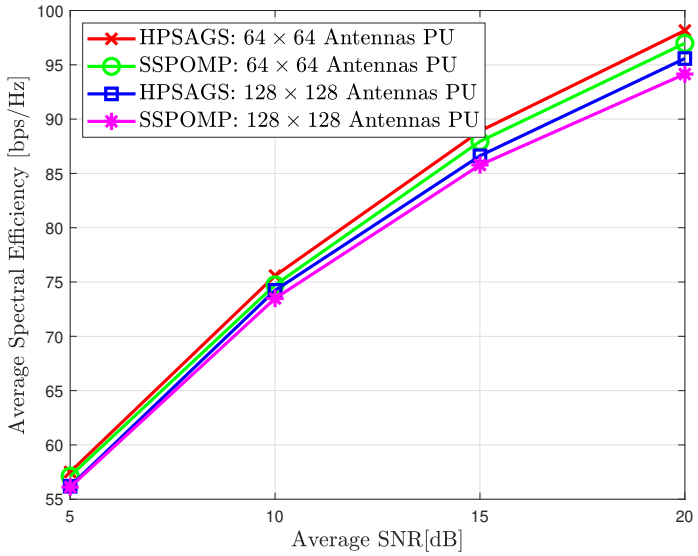


Figure 5.5 64×64 secondary link, 64×64 and 128×128 primary link.

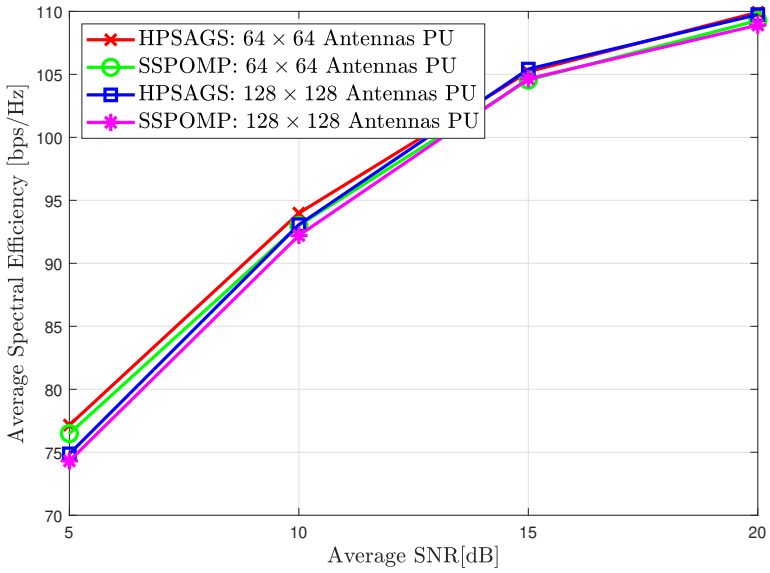


Figure 5.6 128 × 128 secondary link, 64 × 64 and 128 × 128 primary link.

Chapter 6

Summary and Conclusions

The stringent sum-capacity and user data rate requirements of the upcoming 5G cellular networks, in conjunction with the scarcity of resources in the sub-6 GHz spectrum, necessitate the utilization of various techniques to accommodate the projected explosion of mobile data traffic. These include the use of mmWave spectrum, the adoption of spectrum sharing, the densification of the network, and the application of multi-cell MU-MIMO variants based either on coordination / cooperation or on the installation of an excessive number of BS antennas.

Spectrum sharing can provide access to previously reserved for exclusive use bandwidth, thus providing a means for overcoming the crunch of the sub-6 GHz spectrum. Licensed shared access (LSA) represents a paradigm that ensures binary sharing of licensed spectrum, thus providing interference-free operation and predictable QoS to both types of players. However, its conservative nature leaves much to be desired, in terms of the achieved spectrum extension.

Multi-cell MU-MIMO technologies, namely, coordinated multi-point (CoMP) and massive MIMO (mMIMO), can act as enablers of non-orthogonal (or underlay) spectrum sharing. These techniques promise the achievement of substantial SE gains, the provisioning of QoS guarantees to the mobile users, and the protection of the incumbent users from harmful interference, thanks to their advanced interference management and resource allocation features. Therefore, they could get integrated with LSA or its enhancements, namely, dynamic LSA and evolved LSA (eLSA), in a 5G LSA paradigm that provides orthogonal and non-orthogonal access to shared spectrum with QoS guarantees, in order to further extend the usable spectrum, as shown in Fig. 6.1.

Spectrum sharing has been considered also as a spectrum management paradigm for the mmWave segment of the radio spectrum. The motivation behind this has been the lesson learned from the somewhat artificial shortage of resources in the sub-6 GHz spectrum, due to the use of the licensed access spectrum usage model; the high demand for mmWave spectrum licenses not only from MNOs but

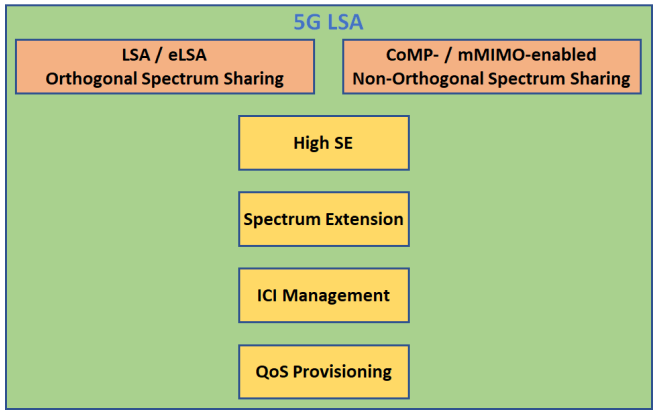


Figure 6.1 The envisioned 5G LSA paradigm.

also from satellite operators, fixed wireless access systems, etc.; and the propagation characteristics in this spectral region (high distance-dependent path loss, high probability of blockage), which facilitate interference management. Due to the high cost and energy consumption of fully digital transceivers operating at mmWave frequencies, a hybrid analog-digital mmWave mMIMO implementation is required. Therefore, efficient hybrid precoding techniques that overcome the limitations of analog processing (i.e., the constant modulus imposed by the phase shifters) should be derived.

In this dissertation, we study the ability of CoMP and mMIMO to enable underlay spectrum sharing with QoS guarantees. In Chapter 2, we study the application of coordinated resource allocation policies for SE maximization under per BS sum-power constraints, per primary receiver interference power constraints, and per mobile user minimum rate constraints for various PS and SS setups under an extensive set of cluster-level and system-level simulations. More specifically, assuming the application of simple and robust standard coordinated linear precoding schemes, such as ZF and regularized ZF precoding, which could accelerate the adoption of this paradigm by commercial deployments, we derive the optimal coordinated power allocation strategies. We note that this approach can achieve significant sum-SE and meet relatively stringent QoS requirements for a somewhat relaxed interference power threshold (IPT) and relatively small inter-system interference received at the MSs. For hard IPT values, we propose a coordinated projected ZF precoding method that improves substantially the performance. We also derive simple suboptimal power allocation methods. The performance improves when the number of BS antennas N increases while the number of users per cell K remains fixed as well as when the number of users per cell increases while the number of BS antennas remains fixed (provided that $K \leq N$), although in the latter case the improvement is not as high as expected due to the reduced

spatial DoF per user and the added inter-user interference by the additional users. The performance improves also when RZF is applied instead of ZF. On the other hand, the sum-SE is reduced when the number of antennas at the primary receiver or the number of primary receivers increases. In Chapter 2 we develop also heuristic coordinated user selection schemes, since the optimal coordinated user selection strategy that is based on exhaustive search over a multi-user multi-cell search space has prohibitively high computational complexity. The proposed reduced search space (RSS) scheme performs close to the optimal strategy, even for moderate search space sizes. The performance degrades slightly for smaller search space sizes or when a greedy implementation is utilized. In any case, RSS scheduling outperforms the other proposed method, namely, inter-system correlation aware user scheduling. Finally, we study the system-level performance, assuming a multi-cluster small-cells setup, since underlay spectrum sharing is better suited in short-range communication applications as previous studies have shown, and taking into account both large-scale and small-scale propagation phenomena. The parameters of the simulation are based on the non-LOS macro-cellular model of 3GPPP. The performance of a proposed dynamic cell clustering scheme is compared with the one achieved with fixed predetermined clusters. We note the same performance trends as in the cluster-level simulations, together with a small performance loss attributed to large-scale fading and OOC interference. We also see that the proposed DCC scheme outperforms the fixed clustering scheme.

The joint transmission (JT) variant of CoMP improves the QoS of the cell-edge users. However, it is rarely applied in practice, due to its stringent demands in terms of fronthaul / backhaul capacity attributed to the requirement for user data sharing among the cooperating BSs. In Chapter 3, we describe a family of coordinated content caching with redundancy enhancement (C3RE) strategies that create JT opportunities to address this issue. We also propose two statistics-based caching schemes, namely, score-gated least recently used (SG-LRU) and SG-Clock (SG-C). An extensive set of simulations reveals that the statistics-based strategies create a significant number of JT opportunities for moderate and large values of the Zipf exponent β, the cache size C, and the sliding window size W. Moreover, they achieve higher local hit rate (LHR) and total hit rate (THR) than LRU, with the best performance obtained for larger values of β, C, and W. Also, the simulations indicate that the C3RE variants that utilize global statistics improve the global hit rate (GHR). We should note that the caching methods that achieve high LHR are preferable over caching schemes that achieve high GHR and have comparable THR with them, since local caching results in higher delay reduction and network traffic savings. We should also mention that the proposed caching schemes not only outperform the de-facto standard LRU method in terms of the achieved cache hit rate, but they maintain its $\mathcal{O}(1)$ cache update effort per request and undercut the loading rate of objects into the cache in case of a cache miss as well. Furthermore, these methods adapt to the dynamics of content popularity,

approach the optimum least frequently used (LFU) hit rate under independent reference model (IRM) conditions, can balance between LRU and LFU by adjusting a single parameter (the size of the sliding window of past requests W), and have the flexibility to use different score-gate functions whose goal might be the optimization of a performance metric other than the cache hit rate. Finally, we propose a hybrid approach where cache-aided JT takes place whenever possible and coordinated precoding (CP) is utilized otherwise and we compare its performance vs. a CP-only scheme, assuming in both cases the use of coordinated ZF precoding and the utilization of a proposed coordinated interference-constrained equal power allocation method. We observe that the hybrid approach outperforms slightly coordinated precoding for relaxed IPT.

In Chapter 4 we propose a method for applying coordinated hybrid ZF precoding to BSs equipped with load-controlled parasitic antenna arrays, as a means to improve the performance for a given number of RF units thanks to the higher array gain. However, in this non-massive regime, the sum-SE enhancement is small, as expected. In order to overcome load computation / load implementation issues, we propose also a beam selection and precoding (BSP) method, where beam selection is based on either CSI feedback or on SINR feedback, as a low overhead alternative. The simulations indicate that there is a high performance loss with this approach – and it gets higher if beam selection is based on SINR feedback instead of CSI feedback. Finally, we describe a coordinated constructive interference zero forcing (CIZF) method. The numerical simulations show that this symbol-level counterpart of coordinated ZF precoding improves the performance in the low SNR regime.

Finally, in Chapter 5 we study the performance of a mmWave mMIMO link that coexists with another such link in an underlay spectrum sharing setup. A computationally efficient hybrid precoding via stochastic approximation with Gaussian smoothing (HPSAGS) method is derived for the determination of the digital and analog precoders and combiners and is compared with unconstrained SVD precoding / combining applied at a fully digital system as well as with spatially sparse precoding with orthogonal matching pursuit (SSPOMP), which is considered the state-of-the-art for hybrid analog-digital mmWave mMIMO links. The applied power allocation schemes have been derived by Zhang et *al*. for coexisting non-massive MIMO links in sub-6 GHz spectrum and can be viewed as special cases of the coordinated power allocation strategies described in Chapter 2. The numerical simulations reveal that the proposed approach outperforms SSPOMP and approaches very closely the unconstrained solution for moderate numbers of transmitted data streams. The performance of HPSAGS degrades slightly for higher number of transmitted data streams, harder IPT values, or primary links with more antennas, and improves when the number of antennas in the secondary link increases. We should note also that in contrast to SSPOMP, HPSAGS can be applied in both sparse (mmWave) and rich (Rayleigh) scattering environments.

We hope that this study will trigger a number of works regarding the efficient use of sub-6 GHz and mmWave spectrum in 5G via the exploitation of contemporary and envisioned multi-cell MU-MIMO technologies. In the future, we plan to extend the work presented in Chapter 2 for multi-antenna MSs and the work presented in Chapter 5 for multi-user mmWave mMIMO setups.

Bibliography

[1] S. Mumtaz, J. Rodriguez, and L. Dai, Eds., *mmWave Massive MIMO: A Paradigm for 5G*. Academic Press, 2017.

[2] "IMT Vision - Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond," ITU, Recommendation ITU-R M.2083-0, September 2015.

[3] "Ericsson Mobility Report," Ericsson, Tech. Rep., June 2019.

[4] "IMT Traffic Estimates for the Years 2020 to 2030," ITU, Report ITU-R M.2370-0, July 2015.

[5] E. Björnson and E. Jorswieck, "Optimal resource allocation in coordinated multi-cell systems," *Foundations and Trends in Communications and Information Theory*, vol. 9, no. 2-3, pp. 113–381, Jan. 2013.

[6] A. Ghosh, J. Zhang, J. G. Andrews, and R. Muhamed, *Fundamentals of LTE*. Prentice Hall, 2011.

[7] Small Cell Forum. [Online]. Available: http://www.smallcellforum.org/

[8] H. Huang, C. B. Papadias, and S. Venkatesan, Eds., *MIMO Communication for Cellular Networks*. Springer, 2012.

[9] "Novel spectrum usage paradigms for 5G," IEEE SIG Cognitive Radio in 5G, White Paper, Nov. 2014.

[10] ICT Regulation Toolkit. [Online]. Available: http://www.ictregulationtoolkit.org/index

[11] J. Restrepo, "ITU-R basics," ITU Regional Radiocommunication Seminar for Americas (RRS-13-Americas), Asunción, Paraguay, 8-12 July 2013. [Online]. Available: https://www.itu.int/en/ITU-R/seminars/rrs/Documents/Intro/IUT-R-Basics.pdf

[12] A. F. Molisch, *Wireless Communications*, 2nd ed. John Wiley & Sons, 2011.

[13] M. D. Mueck, S. Srikanteswara, and B. Badic, "Spectrum Sharing: Licensed Shared Access (LSA) and Spectrum Access System (SAS)," Intel, White Paper, Oct. 2015.

[14] "Spectrum Policy: Analysis of technology trends, future needs and demand for spectrum – Final Report," Analysis Mason, Tech. Rep., 2013.

[15] The digital dividend: Opportunities and challenges. [Online]. Available: https://www.itu.int/net/itunews/issues/2010/01/27.aspx

[16] T. S. Rappaport, R. W. Heath, R. C. Daniels, and J. N. Murdock, *Millimeter Wave Wireless Communications*. Prentice Hall, 2014.

[17] J. H. Schiller, *Mobile Communications*, 2nd ed. Addison-Wesley, 2003.

[18] K. Pahlavan and P. Krishnamurty, *Principles of Wireless Networks: A Unified Approach*. Prentice Hall, 2002.

[19] J. A. Richards, *Radio Wave Propagation: An Introduction for the Non-Specialist*. Springer, 2008.

[20] T. S. Rappaport, *Wireless Communications: Principles and Practice*, 2nd ed. Prentice Hall, 2002.

[21] R. W. Heath, *Introduction to Wireless Digital Communication: A Signal Processing Perspective*. Prentice Hall, 2017.

[22] B. Sklar, "Rayleigh fading channels in mobile digital communication systems part I: Characterization," *IEEE Commun. Mag.*, vol. 35, no. 7, pp. 90–100, July 1997.

[23] A. Goldsmith, *Wireless Communications*. Cambridge University Press, 2005.

[24] Proakis, J. G. and Salehi, M., *Digital Communications*, 5th ed. McGraw-Hill, 2008.

[25] Tse, D. and Viswanath, P., *Fundamentals of Wireless Communication*. Cambridge University Press, 2005.

[26] "Further advancements for E-UTRA physical layer aspects (Release 9)," 2010, 3GPP TS 36.814.

[27] E. Björnson, J. Hoydis, and L. Sanguinetti, "Massive MIMO networks: Spectral, energy, and hardware efficiency," *Foundations and Trends in Signal Processing*, vol. 11, no. 3-4, pp. 154–655, Nov. 2017.

[28] C. D. Nwankwo, L. Zhang, A. Quddus, M. A. Imran, and R. Tafazolli, "A survey of self-interference management techniques for single frequency full duplex systems," *IEEE Access*, vol. 6, pp. 30 242–30 268, Nov. 2017.

[29] J.-G. Remi and C. Letamendia, *LTE Services*. John Wiley & Sons, 2014.

[30] Massive MIMO. [Online]. Available: http://www.massive-mimo.net/

[31] T. L. Marzetta, E. G. Larsson, H. Yang, and H. Q. Ngo, *Fundamentals of Massive MIMO*. Cambridge University Press, 2016.

[32] P. Marsch and G. Fettweis, Eds., *Coordinated Multi-Point in Mobile Communications: From Theory to Practice*. Cambridge University Press, 2011.

[33] M. Ding and L. Hanwen, *Multi-point Cooperative Communication Systems: Theory and Applications*. Springer, 2013.

[34] Q. C. Li, H. Niu, and A. T. Papathanassiou, "5G network capacity: Key elements and technologies," *IEEE Veh. Technol. Mag.*, vol. 9, no. 1, pp. 71–78, Mar. 2014.

[35] R. H. Tehrani, S. Vahid, D. Triantafyllopoulou, H. Lee, and K. Moessner, "Licensed spectrum sharing schemes for mobile operators: A survey and outlook," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 4, pp. 2591–2623, Fourthquarter 2016.

[36] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar *et al.*, "Millimeter wave mobile communications for 5G cellular: It will work!" *IEEE Access*, vol. 1, no. 1, pp. 335–349, May 2013.

[37] S. Hur, T. Kim, D. Love, J. Krogmeier, T. Thomas *et al.*, "Millimeter wave beamforming for wireless backhaul and access in small cell networks," *IEEE Trans. Commun.*, vol. 61, no. 10, pp. 4391–4403, Oct. 2013.

[38] A. Ghosh, T. A. Thomas, M. C. Cudak, R. Ratasuk, P. Moorut *et al.*, "Millimeterwave enhanced local area systems: A high-data-rate approach for future wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1152–1163, June 2014.

[39] L. Wei, R. Q. Hu, Y. Qian, and G. Wu, "Key elements to enable millimetre wave communications for 5G wireless systems," *IEEE Wireless Commun.*, vol. 21, no. 6, pp. 136–143, Dec. 2014.

[40] R. W. Heath. (2015) Comparing Massive MIMO at Sub-6 GHz and Millimeter Wave Using Stochastic Geometry. [Online]. Available: http://www.profheath.org/

[41] T. S. Rappaport, Y. Xing, G. R. MacCartney, A. F. Molisch, and E. Mellios, "Overview of millimeter wave communications for fifth-generation (5G) wireless networks - With a focus on propagation models," *IEEE Trans. on Antennas and Propag.*, vol. 65, no. 12, pp. 6213–6230, Dec. 2017.

[42] I. F. Akyildiz, W.-Y. Lee, M. C. Vuran, and S. Mohanty, "A survey on spectrum management in cognitive radio networks," *IEEE Commun. Mag.*, vol. 46, no. 4, pp. 40–48, Apr. 2008.

[43] L. Doyle, *Essentials of Cognitive Radio*. Cambridge University Press, 2009.

[44] S. Pandit and G. Singh, "An overview of spectrum sharing techniques in cognitive radio communication system," *Wireless Networks*, vol. 23, no. 2, pp. 497–518, Feb. 2017.

[45] C. B. Papadias, K. Ntougias, and G. K. Papageorgiou, "The role of antenna arrays in spectrum sharing," in *Spectrum Sharing: The Next Frontier in Wireless Networks*. IEEE Press, 2019, ch. 12, to appear.

[46] "General Survey of Radio Frequency Bands (30 MHz to 3 GHz): Vienna, Virginia, September 1–5," Shared Spectrum Company, Tech. Rep., 2010, version 2.0.

[47] S. Chen, F. Qin, B. Hu, X. Li, Z. Chen *et al.*, *User-Centric Ultra-Dense Networks for 5G* . Springer, 2018.

[48] T. Q. Duong, X. Chu, and H. A. Suraweera, Eds., *Ultra-Dense Networks for 5G and Beyond: Modelling, Analysis, and Applications*. John Wiley & Sons, 2019.

[49] D. Gesbert, M. Kountouris, R. W. Heath, C.-B. Chae, and T. Saizer, "Shifting the MIMO paradigm," *IEEE Signal Process. Mag.*, vol. 24, no. 5, pp. 36–46, Sep. 2007.

[50] B. Clerckx and C. Oestges, *MIMO Wireless Networks*, 2nd ed. Academic Press, 2013.

[51] E. Castaneda, A. Silva, A. Gameiro, and M. Kountouris, "An overview on resource allocation techniques for multi-user MIMO systems," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 1, pp. 239–284, 2017.

[52] D. Gesbert, S. Hanly, H. Huang, S. S. Shitz, O. Simeone *et al.*, "Multi-cell MIMO cooperative networks: A new look at interference," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 9, pp. 1380–1408, Dec. 2010.

[53] "Coordinated Multi-Point Operation for LTE Physical Layer Aspects," 3GPP, Tech. Rep., Sep. 2013, 3rd Gener. Partnership Project (3GPP), TR 36.819 R11 v11.2.0.

[54] S. Bassoy, H. Farooq, M. A. Imran, and A. Imran, "Coordinated multi-point clustering schemes: A survey," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 2, pp. 743–764, Feb. 2017.

[55] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.

[56] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.

[57] A. K. Gupta, J. G. Andrews, and R. W. Heath, "On the feasibility of sharing spectrum licenses in mmWave cellular systems," *IEEE Trans. Commun.*, vol. 64, no. 9, pp. 3981–3995, Sep. 2016.

[58] A. K. Gupta, A. Alkhateeb, J. G. Andrews, and R. W. Heath, "Gains of restricted secondary licensing in millimeter wave cellular systems," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 11, pp. 2935–2950, Nov. 2016.

[59] X. Chen, "Spectral correlation based multi-antenna spectrum sensing technique," in *IEEE Wireless Commun. and Netw. Conf. (WCNC)*, Las Vegas, NV, USA, Apr. 2008, pp. 735–740.

[60] H. Sadeghi and P. Azmi, "A novel primary user detection method for multiple-antenna cognitive radio," in *Int.Symp. Telecommun. (IST)*, Tehran, Iran, Aug. 2008, pp. 188–192.

[61] H. Jitvanichphaibool, Y.-C. Liang, and Y. Zeng, "Spectrum sensing using multiple antennas for spatially and temporally correlated noise environments," in *IEEE Symp. New Frontiers in Dynamic Spectr. Access Netw. (DySPAN)*, Singapore, Apr. 2010, pp. 1–7.

[62] E. P. Tsakalaki, "Spatial spectrum sensing for wireless handheld terminals: design challenges and novel solutions based on tunable parasitic antennas," *IEEE Wireless Commun. Mag.*, vol. 17, no. 4, pp. 33–40, Aug. 2010.

[63] P. Uriza, E. Rebeiz, and D. Cabric, "Multiple antenna cyclostationary spectrum sensing based on the cyclic correlation significance test," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 11, pp. 2185–2195, Nov. 2013.

[64] N. Pratas, N. Marchetti, N. R. Prasad, A. Rodrigues, and R. Prasad, "Centralized cooperative spectrum sensing for ad-hoc disaster relief network clusters," in *Proc. IEEE Int. Conf. on Commun.*, Cape Town, South Africa, May 2010, pp. 1–5.

[65] ——, "Decentralized cooperative spectrum sensing for ad-hoc disaster relief network clusters," in *Proc. IEEE Veh. Technol. Conf. (VTC)*, Taipei, Taiwan, May 2010, pp. 1–5.

[66] N. Pratas, N. R. Prasad, A. Rodrigues, and R. Prasad, "Cooperative spectrum sensing: state of the art review," in *Proc. IEEE Intern. Conf. Wireless Commun. Veh. Technol. Information Theory Aerosp. and Electron. Syst. Technol. (Wireless VITAE)*, Mar. 2011, pp. 1–6.

[67] ——, "Spatial diversity aware data fusion for cooperative spectrum sensing," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, Bucharest, Romania, Aug. 2012, pp. 2669–2673.

[68] "TV White Space," GSMA, Public Policy Position, Feb. 2016.

[69] R. Ramjee, S. Roy, and K. Chintalapudi, "A critique of FCC's TV White Space regulations," *GetMobile: Mobile Computing and Communications*, vol. 20, no. 1, pp. 20–25, Jan. 2016.

[70] A. Zavodovski, "Understanding the challenges of TV White Space databases for mobile usage," Master's thesis, University of Helsinki, May 2016.

[71] "ETSI TR 103 113 V1.1.1: Mobile Broadband Services in the 2300 MHz – 2400 MHz Frequency Band under Licensed Shared Access Regime," ETSI Technical Committee Reconfigurable Radio Systems (TC RRS), System Reference Document, July 2013.

[72] "ECC Report 205: Licensed Shared Access," ECC WG FM53, Tech. Rep., Feb. 2014.

[73] "ECC Decision (14)02: Harmonised Technical and Regulatory Conditions for the Use of the Band 2300-2400 MHz for Mobile/Fixed Communications Networks (MFCN)," ECC, Tech. Rep., June 2014.

[74] "ETSI TS 103 154 V1.1.1: System Requirements for Operation of Mobile Broadband Systems in the 2300 MHz - 2400 MHz Band under Licensed Shared Access (LSA)," ETSI Technical Committee Reconfigurable Radio Systems (TC RRS), Technical Specification, Oct. 2014.

[75] "ETSI TS 103 235 V1.1.1: System Architecture and High Level Procedures for Operation of Licensed Shared Access (LSA) in the 2300 MHz - 2400 MHz Band," ETSI Technical Committee Reconfigurable Radio Systems (TC RRS), Technical Specification, Oct. 2015.

[76] "ETSI TS 103 379 V1.1.1: Information Elements and Protocols for the Interface Between LSA Controller (LC) and LSA Repository (LR) for Operation of Licensed Shared Access (LSA) in the 2300 MHz - 2400 MHz Band," ETSI Technical Committee Reconfigurable Radio Systems (TC RRS), Technical Specification, Jan. 2017.

[77] "Authorised Shared Access - An evolutionary spectrum authorisation scheme for sustainable economic growth and consumer benefit," Qualcomm, NSN, Input Document FM(11)116 to the 72nd Meeting of the WG FM, May 2011.

[78] "FM(12)084 Annex 47: Report on ASA concept," CEPT/ECC, Tech. Rep., 2012.

[79] A. Morgado, A. Gomes, V. Frascolla, K. Ntougias, C. B. Papadias *et al.*, "Dynamic LSA for 5G networks: The ADEL perspective," in *Proc. Eur. Conf. Netw. Commun. (EuCNC)*, Paris, France, 2015, pp. 190–194.

[80] "ETSI TR 103 588 V1.1.1: Feasibility Study on Temporary Spectrum Access for Local High-Quality Wireless Networks," ETSI Technical Committee Reconfigurable Radio Systems (TC RRS), Tech. Rep., Feb. 2018.

[81] "FCC 15-47: REPORT AND ORDER AND SECOND FURTHER NOTICE OF PROPOSED RULEMAKING," FCC, In the Matter of Amendment of the Commission's Rules with Regard to Commercial Operations in the 3550-3650 MHz Band, Apr. 2015.

[82] D. Denkovski, V. Rakovic, V. Atanasovski, L. Gavrilovska, and O. Mahonen, "Generic multiuser coordinated beamforming for underlay spectrum sharing," *IEEE Trans. Commun.*, vol. 64, no. 6, pp. 2285–2298, June 2016.

[83] L. Wang, H. Q. Ngo, M. Elkashlan, Q. Duong, and K.-K. Wong, "Massive MIMO in spectrum sharing networks: Achievable rate and power efficiency," *IEEE Systems Journal*, vol. 11, no. 1, pp. 20–31, Mar. 2017.

[84] P. K. Agyapong, M. Iwamure, D. Staehle, W. Kiess, and A. Benjebbour, "Design considerations for a 5G network architecture," *IEEE commun. mag.*, vol. 52, no. 11, pp. 65–75, Nov. 2014.

[85] A. Mohamed, O. Onireti, M. A. Imran, A. Imran, and R. Tafazzoli, "Control-data separation architecture for cellular radio access networks: A survey and outlook," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 1, pp. 446–465, 1st Quart. 2015.

[86] J. G. Andrews, X. Zhang, G. D. Durgin, and A. K. Gupta, "Are we approaching the fundamental limits of wireless network densification?" *IEEE Commun. Mag.*, vol. 54, no. 10, pp. 184–190, Oct. 2016.

[87] J. Wu, Z. Zhang, Y. Hong, and Y. Wenn, "Cloud radio access network (C-RAN): A primer," *IEEE Netw.*, vol. 29, no. 1, pp. 35–41, Jan.–Feb. 2015.

[88] G. Hasslinger and K. Ntougias, "Evaluation of caching strategies based on access statistics of past requests," in *Int. GI/ITG MMB and DFT Conference*, ser. Lecture Notes in Computer Science (LNCS), Springer, Ed., vol. 8376, Bamberg, Germany, Mar. 2014, pp. 120–135.

[89] G. Hasslinger, K. Ntougias, and F. Hasslinger, "A new class of web caching strategies for content delivery," in *Int. Telecommun. Netw. Strategy and Planning Symp. (Networks)*, Funchal, Portugal, 2014, 17-19 Sept.

[90] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: Wireless video content delivery through distributed caching helpers," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 8402–8413, Dec. 2013.

[91] F. Zhou, L. Fan, N. Wang, G. Luo, J. Tang *et al.*, "A cache-aided communication scheme for downlink coordinated multipoint transmission," *IEEE Access*, vol. 6, pp. 1416–1427, Dec. 2017.

[92] I. Garcia, N. Kusashima, K. Sakaguchi, and K. Araki, "Dynamic cooperation set clustering on base station cooperation cellular networks," in *Proc. IEEE Intern. Symp. Pers. Indoor and Mobile Radio Commun. (PIMRC)*, 2010, pp. 2127–2132.

[93] P. Marsch and G. Fettweis, "Static clustering for cooperative multi-point (CoMP) in mobile communications," in *Proc. IEEE Intern. Conf. Commun. (ICC)*, 2011.

[94] A. Papadogiannis, D. Gesbert, and E. Hardouin, "A dynamic clustering approach in wireless networks with multi-cell cooperative processing," in *Proc. IEEE Intern. Conf. Commun. (ICC)*, 2008.

[95] E. Björnson, N. Jalden, M. Bengtsson, and B. Ottersten, "Optimality properties, distributed strategies, and measurement-based evaluation of coordinated multicell OFDMA transmission," *IEEE Trans. Signal Process.*, vol. 59, no. 12, pp. 6086–6101, Dec. 2011.

[96] H. Hu, G. Caire, H. Papadopoulos, and S. Ramprashad, "Achieving massive MIMO spectral efficiency with a not-so-large number of antennas," *IEEE Trans. Wireless Commun.*, vol. 11, no. 9, pp. 3226–3239, Sep. 2012.

[97] S. Kaviani and W. Kryzmien, "Multicell scheduling in network MIMO," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, 2010.

[98] S. Kaviani, O. Simeone, W. Krzymien, and S. Shamai, "Linear precoding and equalization for network MIMO with partial cooperation," *IEEE Trans. Veh. Technol. (VTC)*, vol. 61, no. 5, pp. 2083–2095, June 2012.

[99] A. Tolli, M. Codreanu, and M. Juntti, "Cooperative MIMO-OFDM cellular system with soft handover between distributed base station antennas," *IEEE Trans. Wireless Commun.*, vol. 7, no. 4, pp. 1428–1440, Apr. 2008.

[100] A. Kalis, A. G. Kanatas, and C. B. Papadias, Eds., *Parasitic Antenna Arrays for Wireless MIMO Systems*. Springer, 2014.

[101] A. F. Molisch, V. V. Ratnam, S. Han, Z. Li, S. L.-H. Nguyen *et al.*, "Hybrid beamforming for massive MIMO–A survey," *IEEE Commun. Mag.*, vol. 55, no. 9, pp. 134–141, Sep. 2017.

[102] O. El Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1499–1513, Mar. 2014.

[103] A. Alkhateeb, G. Leus, and R. W. Heath, "Limited feedback hybrid precoding for multi-user millimeter wave systems," *IEEE Trans. Wireless Commun.*, vol. 14, no. 11, pp. 6481–6494, Nov. 2015.

[104] M. Iwanow, N. Vucic, M. H. Castaneda, J. Luo, W. Xu *et al.*, "Some aspects on hybrid wideband transceiver design for mmWave communication systems," in *20th International ITG Workshop on Smart Antennas (WSA)*, Munich, Germany, 9-11 Mar. 2016, pp. 1–8.

[105] A. Tuholukova, G. Neglia, and T. Spyropoulos, "Optimal cache allocation for femto helpers with joint transmission capabilities," in *IEEE Int. Conf. Communications (ICC)*, Paris, France, May 2017.

[106] W. C. Ao and K. Psounis, "Fast content delivery via distributed caching and small cell cooperation," *IEEE Trans. Mobile Comput.*, vol. 17, no. 5, pp. 1048–1061, May 2017.

[107] 3GPP, "3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Further advancements for E-UTRA physical layer aspects (Release 9)," Technical Report, Mar. 2010, 3GPP TR 36.814.

[108] 5GPPP, "5G Vision," White Paper, 2015.

[109] R. Zhang and Y. C. Liang, "Exploiting multi-antennas for opportunistic spectrum sharing in cognitive radio networks," *IEEE J. Sel. Topics in Signal Process.*, vol. 2, no. 1, pp. 88–102, Feb. 2008.

[110] I. Turki, I. Kammoun, and M. Siala, "Beamforming design and sum rate maximization for the downlink of underlay cognitive radio networks," in *Int. Wireless Commun. and Mobile Comput. Conf. (IWCMC)*, Dubrovnik, Croatia, 2015, pp. 178–183, 24-28 Aug. 2015.

[111] V. D. Nguyen, L.-N. Tran, T. Q. Duong, O.-S. Shin, and R. Farell, "An efficient precoder design for multiuser MIMO cognitive radio networks with interference constraints," *IEEE Trans. Veh. Technol.*, vol. 66, no. 5, pp. 3991–4004, May 2017.

[112] L. Claudino and T. Abrao, "Efficient ZF-WF strategy for sum-rate maximization of MU-MISO cognitive radio networks," *AEU Int. J. Electronics Commun.*, vol. 84, pp. 366–374, Feb. 2018.

[113] L. Gallo, F. Negro, I. Ghauri, and D. T. M. Slock, "Weighted sum rate maximization in the underlay cognitive MISO interference channel," in *22nd Int. IEEE Symp. on Pers., Indoor and Mobile Radio Commun. (PIMRC)*, Toronto, ON, Canada, 2011, pp. 661–665, 11-14 September.

[114] K. Ntougias, G. K. Papageorgiou, C. B. Papadias, T. B. Sorensen, and M. Sellathurai, "Low-complexity coordinated resource allocation for QoS-constrained SR maximization in underlay specrum sharing setups," *IEEE Trans. Wireless Commun.*, 2019, submitted.

[115] K. Ntougias, D. Ntaikos, C. B. Papadias, and G. K. Papageorgiou, "Simple cooperative transmission schemes for underlay spectrum sharing using symbol-level precoding and load-controlled arrays," in *Int. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, Brighton, UK, 2019, 12-17 May.

[116] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[117] A. Sani, M. M. Feghhi, and A. Abbasfar, "Discrete bit loading and power allocation for OFDMA downlink with minimum rate guarantee for users," *AEU - Int. J. Electronics and Commun.*, vol. 68, no. 7, pp. 602–610, 2014, July.

[118] J. Mirza, P. J. Smith, and P. A. Dmochowski, "Coordinated regularized zero-forcing precoding for multicell MISO systems with limited feedback," *IEEE Trans. Veh. Technol.*, vol. 66, no. 1, pp. 335–343, Jan. 2017.

[119] G. Hasslinger, K. Ntougias, F. Hasslinger, and O. Hohlfeld, "Performance evaluation for new web caching strategies combining LRU with score based object selection," *Computer Networks*, vol. 125, pp. 172–186, Oct. 2017.

[120] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "I Tube, You Tube, Everybody Tubes: Analyzing the world's largest user generated content video system," in *ACM SIGCOMM Conf. Internet Measurement (IMC)*, San Diego, CA, USA, 2007, pp. 1–14, 24-26 October.

[121] T. Qiu, Z. Ge, S. Lee, J. Wang, Q. Zhao *et al.*, "Modeling channel popularity dynamics in a large IPTV system," in *SIGMETRICS Int. Joint Conf. on*

Measurement and Modeling of Computer Systems, Seattle, WA, USA, 2009, pp. 275–286, June.

[122] K. Ntougias, C. B. Papadias, G. K. Papageorgiou, and G. Hasslinger, "Spectral coexistence of 5G networks and satellite communication systems enabled by coordinated caching and QoS-aware resource allocation," in *Eur. Signal Process. Conf. (EUSIPCO)*, A Coruna, Spain, Sep. 2019, to appear.

[123] K. Ntougias, C. B. Papadias, G. K. Papageorgiou, G. Hasslinger, and T. B. Sorensen, "Coordinated caching and QoS-aware resource allocation for spectrum sharing," *Wireless Pers. Commun.*, 2019, to appear.

[124] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and zipf-like distributions: Evidence and implications," in *18th IEEE Infocom*, New York, NY, USA, 1999, pp. 126–134, 21-25 March.

[125] R. Bolla, R. Gaetta, A. Magnetto, and M. Sciuto, "A measurement study supporting P2P file-sharing community models," *Computer Networks*, vol. 53, no. 4, pp. 485–500, Mar. 2009.

[126] C. Fricker, P. Robert, and J. Roberts, "A versatile and accurate approximation for LRU cache performance," in *Int. Teletraffic Congr. (ITC)*, Krakow, Poland, 2012.

[127] G. Hasslinger, K. Ntougias, F. Hasslinger, and O. Hohlfeld, "Comparing web cache implementations for fast O(1) updates based on LRU, LFU and score gated strategies," in *IEEE Int. Workshop on Computer Aided Model. and Design of Commun. Links and Netw. (CAMAD)*, Barcelona, Spain, 2018, 17-19 Sept.

[128] K. Ntougias, D. Ntaikos, C. B. Papadias, and G. K. Papageorgiou, "Coordinated hybrid precoding and QoS-aware power allocation for underlay spectrum sharing with load-controlled antenna arrays," in *IEEE Signal Process. Advances Wireless Commun. (SPAWC)*, Cannes, France, 2019, 2-5 July.

[129] I. Ahmed, H. Khammari, A. Shahid, A. Musa, K. S. Kim *et al.*, "A survey on hybrid beamforming techniques in 5G: Architecture and system model perspectives," *IEEE Commun. Surveys & Tutorials*, vol. 20, no. 4, pp. 3060–3097, 2018, fourthquarter.

[130] K. Ntougias, D. Ntaikos, and C. B. Papadias, "Robust low-complexity arbitrary user- and symbol-level multi-cell precoding with single-fed load-controlled parasitic antenna arrays," in *Int. Conf. Telecommun. (ICT)*, Thessaloniki, Greece, 2016, 16-18 May.

[131] ——, "Coordinated MIMO with single-fed load-controlled parasitic antenna arrays," in *IEEE Signal Process. Advances Wireless Commun. (SPAWC)*, Edinburgh, UK, 3-6 July 2016.

[132] K. Ntougias, D. Ntaikos, B. Gizas, G. K. Papageorgiou, and C. B. Papadias, "Large load-controlled multiple-active multiple-passive antenna arrays: Transmit beamforming and multi-user precoding," in *Eur. Signal Process. Conf. (EUSIPCO)*, Kos, Greece, 2017, 28 Aug.-2 Sept. 2017.

[133] K. Ntougias, D. Ntaikos, and C. B. Papadias, "Single- and multiple-RF load-controlled parasitic antenna arrays operating at cm-wave frequencies: Design and application for 5G wireless access / backhaul," in *FITCE Congress*, Madrid, Spain, 2017, Sept. 14-15.

[134] G. Alexandropoulos *et al.*, "Precoding for multiuser MIMO systems with single-fed parasitic antenna arrays," in *IEEE Global Commun. Conf. (GLOBECOM)*, Austin, TX, USA, Dec. 2014, pp. 3897–3902.

[135] V. I. Barousis and C. B. Papadias, "Arbitrary precoding with single-fed parasitic arrays: Closed-form expressions and design guidelines," *IEEE Wireless Commun. Letters*, vol. 3, no. 2, pp. 229–232, Feb. 2014.

[136] B. Han, V. I. Barousis, C. B. Papadias, A. Kalis, and R. Prasad, "MIMO over ESPAR with 16-QAM modulation," *IEEE Wireless Commun. Letters*, vol. 2, no. 6, pp. 687–690, Dec. 2013.

[137] L. Zhou, F. A. Khan, T. Ratnarajah, and C. B. Papadias, "Achieving arbitrary signals transmission using a single radio frequency chain," *IEEE Trans. Commun.*, vol. 63, no. 12, pp. 4865–4878, Oct. 2015.

[138] A. Li, C. Masouros, and C. B. Papadias, "MIMO Transmission for single-fed ESPAR with quantized loads," *IEEE Trans. Commun.*, vol. 65, no. 7, pp. 2863–2876, 2017, July.

[139] C. Masouros and E. Alsusa, "Dynamic linear precoding for the exploitation of known interference in MIMO broadcast systems," *IEEE Trans. Wireless Commun.*, vol. 8, no. 3, pp. 1396–1404, Mar. 2009.

[140] M. A. Styblinski and T.-S. Tang, "Experiments in nonconvex optimization: Stochastic approximation with function smoothing and simulated annealing," *Neural Netw.*, vol. 3, no. 4, pp. 467–483, 1990.

[141] G. K. Papageorgiou, M. Sellathurai, K. Ntougias, and C. B. Papadias, "A stochastic optimization approach to hybrid precoding in massive MIMO systems," *IEEE Wireless Commun. Letters*, 2019, submitted.

[142] A. K. Gupta and D. K. Nagar, *Matrix Variate Distributions*. Chapman and Hall/CRC, 2018.

[143] http://github.com/ge99210/Hybrid-Precoding-Combining-, [Online; accessed 19 July 2019].

Appendix A

Proof of Theorem 2.1

Let us prove Theorem 2.1, considering initially the scenario where the PS is either a SISO link or a MIMO one [114]. Then, the Lagrangian form of **P1** is:

$$\begin{aligned} L\left(P_{mk}^{m}, \nu_m, \mu, \xi_{mk}^{m}\right) = & -\sum_{m=1}^{M}\sum_{k=1}^{K}\log_2\left(1+\lambda_{mk}^{m}P_{mk}^{m}\right) \\ & +\sum_{m=1}^{M}\nu_m\left(\sum_{k=1}^{K}P_{mk}^{m}-P_T\right) \\ & +\mu\left(\sum_{m=1}^{M}\sum_{k=1}^{K}\alpha_{mk}^{m}P_{mk}^{m}-P_I\right) \\ & -\sum_{m=1}^{M}\sum_{k=1}^{K}\xi_{mk}^{m}\left(P_{mk}^{m}-\tilde{P}_{mk}\right), \end{aligned} \tag{A.1}$$

where ν_m, μ and ξ_{mk}^{m} are the Lagrange multipliers that are associated with the SPCs, the IPC, and the QoS constraints.

The Karush-Kuhn-Tucker (KKT) (or optimality) conditions [116] are:

$$-\frac{\lambda_{mk}^{m}}{\ln(2)\left(1+\lambda_{mk}^{m}P_{mk}^{m}\right)}+\nu_m+\mu a_{mk}^{m}-\xi_{mk}^{m}=0. \tag{A.2a}$$

$$\xi_{mk}^{m}\geq 0. \tag{A.2b}$$

$$\xi_{mk}^{m}\left(P_{mk}^{m}-\tilde{P}_{mk}\right)=0. \tag{A.2c}$$

$$\nu_m\geq 0. \tag{A.2d}$$

$$\nu_m\left(\sum_{k=1}^{K}P_{mk}^{m}-P_T\right)=0. \tag{A.2e}$$

$$\mu\geq 0. \tag{A.2f}$$

$$\mu\left(\sum_{m=1}^{M}\sum_{k=1}^{K}\alpha_{mk}^{m}P_{mk}^{m}-P_I\right)=0. \tag{A.2g}$$

From Eq. (A.2a) and Eq. (A.2c) and using the slack variable elimination technique we get:

$$\left(\nu_m+\mu\alpha_{mk}^{m}-\frac{\lambda_{mk}^{m}}{\ln(2)\left(1+\lambda_{mk}^{m}P_{mk}^{m}\right)}\right)\left(P_{mk}^{m}-\tilde{P}_{mk}\right)=0. \tag{A.3}$$

Moreover, from Eq. (A.2b) we obtain:

$$\nu_m+\mu\alpha_{mk}^{m}\geq\frac{\lambda_{mk}^{m}}{\ln(2)\left(1+\lambda_{mk}^{m}P_{mk}^{m}\right)}. \tag{A.4}$$

Thus, we distinguish two cases: If $\nu_m+\mu\alpha_{mk}^{m}<\lambda_{mk}^{m}/\ln(2)\left(1+\lambda_{mk}^{m}\tilde{P}_{mk}\right)$, then $P_{mk}^{m}>\tilde{P}_{mk}$ and from Eq. (A.3):

$$P_{mk}^{m}=\frac{1}{\ln(2)\left(\nu_m+\mu\alpha_{mk}^{m}\right)}-\frac{1}{\lambda_{mk}^{m}}. \tag{A.5}$$

Otherwise, we realize from Eq. (A.3) that $P_{mk}^{m}=\tilde{P}_{mk}$. Therefore, the solution to **P1** is given by Eq. (2.33).

The Lagrangian form and KKT conditions of **P2** are derived from those of **P1** if we set $\tilde{P}_{mk}=0$. Thus, we obtain from Eq. (A.2a) and Eq. (A.2c):

$$\left(\nu_m+\mu\alpha_{mk}^{m}-\frac{\lambda_{mk}^{m}}{\ln(2)\left(1+\lambda_{mk}^{m}P_{mk}^{m}\right)}\right)P_{mk}^{m}=0. \tag{A.6}$$

Then, based on Eq. (A.4), we distinguish two cases: If $\nu_m+\mu a_{mk}^{m}<\lambda_{mk}^{m}/\ln(2)$, then

$P_{mk}^m > 0$ and from Eq. (A.6):

$$P_{mk}^m = \frac{1}{\ln(2)\left(\nu_m + \mu\alpha_{mk}^m\right)} - \frac{1}{\lambda_{mk}^m}. \tag{A.7}$$

Otherwise (i.e., if $\nu_m + \mu\alpha_{mk}^m \geq \lambda_{mk}^m/\ln(2)$), we realize from Eq. (A.6) that $P_{mk}^m = 0$. Therefore, the solution to **P2** is given by Eq. (2.34).

Similarly, the Lagrangian form and KKT conditions of **P3** are derived from those of **P2** if we set $\mu = 0$. Thus, we obtain from Eq. (A.2a) and Eq. (A.2c):

$$\left(\nu_m - \frac{\lambda_{mk}^m}{\ln(2)\left(1 + \lambda_{mk}^m P_{mk}^m\right)}\right) P_{mk}^m = 0. \tag{A.8}$$

Then, based on Eq. (A.4), we distinguish two cases: If $\nu_m < \lambda_{mk}^m/\ln(2)$, then $P_{mk}^m > 0$ and from Eq. (A.8):

$$P_{mk}^m = \frac{1}{\ln(2)\nu_m} - \frac{1}{\lambda_{mk}^m}. \tag{A.9}$$

Otherwise (i.e., if $\nu_m \geq \lambda_{mk}^m/\ln(2)$), we realize from Eq. (A.8) that $P_{mk}^m = 0$. Therefore, the solution to **P3** is given by Eq. (2.35).

Finally, the Lagrangian form and KKT conditions of **P4** are derived from those of **P1** if we set $\mu = 0$. Thus, we obtain from Eq. (A.2a) and Eq. (A.2c):

$$\left(\nu_m - \frac{\lambda_{mk}^m}{\ln(2)\left(1 + \lambda_{mk}^m P_{mk}^m\right)}\right)\left(P_{mk}^m - \tilde{P}_{mk}\right) = 0. \tag{A.10}$$

Moreover, from Eq. (A.2b) we obtain:

$$\nu_m \geq \frac{\lambda_{mk}^m}{\ln(2)\left(1 + \lambda_{mk}^m P_{mk}^m\right)}. \tag{A.11}$$

Thus, we distinguish two cases: If $\nu_m < \lambda_{mk}^m/\ln(2)\left(1 + \lambda_{mk}^m \tilde{P}_{mk}\right)$, then $P_{mk}^m > \tilde{P}_{mk}$ and from Eq. (A.3):

$$P_{mk}^m = \frac{1}{\ln(2)\nu_m} - \frac{1}{\lambda_{mk}^m}. \tag{A.12}$$

Otherwise, we realize from Eq. (A.3) that $P_{mk}^m = \tilde{P}_{mk}$. Therefore, the solution to **P4** is given by Eq. (2.36).

Next, let us consider the case where the PS is a MIMO BC with Q single-antenna primary receivers [114]. Then, the term $\mu\left(\sum_{m=1}^M \sum_{k=1}^K \alpha_{mk}^m P_{mk}^m - P_I\right)$ in Eq. (A.1) is replaced by the term $\left(\sum_{q=1}^Q \sum_{m=1}^M \sum_{k=1}^K \mu_q \left(\alpha_{mk}^m\right)^{(q)} P_{mk}^m - P_I\right)$. Similarly, Eq. (A.2a) and Eq. (A.2g) are replaced by Eq. (A.13a) and Eq. (A.13b), respectively:

$$-\frac{\lambda_{mk}^{m}}{\ln(2)\left(1+\lambda_{mk}^{m}P_{mk}^{m}\right)}+\nu_m+\sum_{q=1}^{Q}\mu_q\left(\alpha_{mk}^{m}\right)^{(q)}-\xi_{mk}^{m}=0. \tag{A.13a}$$

$$\left(\sum_{q=1}^{Q}\sum_{m=1}^{M}\sum_{k=1}^{K}\mu_q\left(\alpha_{mk}^{m}\right)^{(q)}P_{mk}^{m}-P_I\right)=0. \tag{A.13b}$$

Let's set $\left(P_{mk}^{m}-\tilde{P}_{mk}\right)=P_d$. Then, from Eq. (A.13a) and Eq. (A.2c) we obtain:

$$\left(\nu_m+\sum_{q=1}^{Q}\mu_q\left(\alpha_{mk}^{m}\right)^{(q)}-\frac{\lambda_{mk}^{m}}{\ln(2)\left(1+\lambda_{mk}^{m}P_{mk}^{m}\right)}\right)P_d=0. \tag{A.14}$$

Moreover, from Eq. (A.2b) we obtain:

$$\nu_m+\sum_{q=1}^{Q}\mu_q\left(\alpha_{mk}^{m}\right)^{(q)}\geq\frac{\lambda_{mk}^{m}}{\ln(2)\left(1+\lambda_{mk}^{m}P_{mk}^{m}\right)}. \tag{A.15}$$

Thus, we distinguish two cases: If $\nu_m+\sum_{q=1}^{Q}\mu_q\left(\alpha_{mk}^{m}\right)^{(q)}<\lambda_{mk}^{m}/\ln(2)\left(1+\lambda_{mk}^{m}\tilde{P}_{mk}\right)$, then $P_{mk}^{m}>\tilde{P}_{mk}$ and from Eq. (A.14):

$$P_{mk}^{m}=\frac{1}{\ln(2)\left(\nu_m+\sum_{q=1}^{Q}\mu_q\left(\alpha_{mk}^{m}\right)^{(q)}\right)}-\frac{1}{\lambda_{mk}^{m}}. \tag{A.16}$$

Otherwise, we realize from Eq. (A.14) that $P_{mk}^{m}=\tilde{P}_{mk}$. Therefore, the solution to **P1** for this scenario is given by Eq. (2.33) if we replace $\alpha_{mk}^{m}\mu$ by $\sum_{q=1}^{Q}\mu_q\left(\alpha_{mk}^{m}\right)^{(q)}$. It is also straightforward to show that the solution to **P2** for this scenario is given by Eq. (2.34) using the aforementioned substitution of variables.

SUMMARY

The use of multi-cell multi-user MIMO technologies as underlay spectrum sharing enablers promises substantial spectral efficiency gains, extension of the available spectrum, and provision of QoS guarantees. In this dissertation, we derive and comparatively evaluate via an extended set of numerical simulations various resource allocation techniques for coordinated multi-point and massive MIMO setups that coexist with incumbent systems. These include coordinated linear precoding schemes and corresponding optimal coordinated QoS-aware power allocation methods, heuristic and greedy coordinated user selection algorithms, simple yet efficient dynamic cell clustering techniques, coordinated caching strategies and statistic-based caching schemes for cache-aided joint transmission, coordinated codeword-level and symbol-level precoding methods for nodes equipped with load-controlled antenna arrays as well as a beam selection and precoding alternative, and a hybrid precoding / combining method achieved via stochastic approximation with Gaussian smoothing for millimeter-wave massive MIMO links. This study sheds light on the effect of various parameters on the performance of these techniques, provides design guidelines, and paves the way for the efficient use of the spectrum in 5G systems via an integrated multi-cell MIMO / licensed shared access (LSA) paradigm.

ISSN (online): 2446-1628
ISBN (online): 978-87-7210-502-4 AALBORG UNIVERSITETSFORLAG