600 Jahre Universität Rostock

Traditio et Innovatio

Dissertation zum Thema

Centralized and Partial Decentralized Design for the Fog Radio Access Network

zur Erlangung des akademischen Grades
Doktor-Ingenieur (Dr.-Ing.)

Vorgelegt von:	**M.Sc. Di Chen**
Vorsitzender:	Prof. Dr.-Ing. Ralf Salomon (Universität Rostock)
Erstgutachter:	Prof. Dr.-Ing. Volker Kühn (Universität Rostock)
Zweitgutachter:	Prof. Dr.-Ing. Armin Dekorsy (Universität Bremen)
Tag der Einreichung:	27.03.2021
Tag der Verteidigung:	09.07.2021

https://doi.org/10.18453/rosdok_id00003187

Institut für Nachrichtentechnik
Fakultät für Informatik und Elektrotechnik
Universität Rostock

Printed on acid-free and elemental chlorine-free bleached paper.

As of:: September 13, 2021

Abstract

Cloud Radio Access Network (C-RAN) has attracted an explosive enthusiasm among researchers worldwide in recent years. The basic concept for the C-RAN is rather simple and straightforward: Moving as much base band signal processing functionalities as possible to the cloud, in order to achieve a centralized processing and joint optimization. In the uplink, the densely and widely distributed Remote Radio Heads (RRH) positioning on edges of the network perform only rather basic Radio Frequency (RF) functions, which act only as signal collectors without implementing any complicated signal processing steps. The collected signals are then delivered, via the capacity-limited fronthauls, to the Base Band Units (BBU) pool located in the cloud. At the BBU pool, further based band signal processing procedures are executed jointly in a centralized manner. The downlink is similar, the BBU pool executes most base band signal processing steps, as well as some higher layer functionalities, before the data streams are sent to RRHs. Due to such joint and centralized processing in the cloud, much more efficient interference management, resource allocation, traffic handling, etc., can be realized, which can lead to much higher Spectral Efficiency (SE) and Energy Efficiency (EE) of the network. Hence, C-RAN is shown to be a promising network architecture for the Fifth Generation (5G) wireless system. In order to combat against some accompanied emerging drawbacks and practical difficulties of such centralized processing, e,g, high latency, high computational complexity imposed on the BBU pool, and high capacity demand on the fronthauls, etc., the Fog Radio Access Network (F-RAN), based on the fog computing (edge computing), has been proposed and widely discussed recently. In F-RAN, the RRH evolves into the so-called enhanced RRH (eRRH). There are various strategies in the realization of an eRRH in practice. For example, equipping a RRH with some limited computational capabilities, or simply adding a cache module to it. With the fog computing, several selected base band signal processing functionalities can be pulled back from the cloud to the network edge. With such a structure, some shortcomings of C-RAN can be overcome, while many benefits can still be retained. Naturally, compared to C-RAN, some performance degradation is inevitable.

In this work, we investigate the design and optimization for F-RAN. In order to fulfill different requirements for various 5G scenarios, we take different criteria into

consideration, e.g., high Energy Efficiency (EE) oriented design, and high Spectral Efficiency (SE) oriented design. For each architecture, both uplink and downlink are considered. Furthermore, we tackle this problem in two steps: In the first step, we propose the framework of joint optimization and design, where all optimization tasks are performed in a centralized manner at the BBU pool, the global Channel State Information (CSI) is thus required. Therefore, a large amount of overhead has to be conveyed from the network edge to the cloud, which would impair the actual performance of the network. Although the centralized design is theoretical optimal, its computational complexity might be prohibitively high in some cases, and the amount of overhead can also be intolerable. Therefore, we proceed to the second step: With the help of the edge computing, as well as the channel hardening effects from the concept of Massive MIMO, the framework of a partially decentralized signal processing mechanism and optimization are proposed. In this approach, only partial CSI is required at the BBU pool in the cloud. Thus, the amount of overhead can be greatly reduced. Moreover, as we are going to show, the computational complexity, and even the hardware costs can also be reduced.

Besides the assumption of perfect CSI, the robust design and optimization of the network based on inaccurate CSI is also to be investigated. Compared to the conventional network architecture, the imperfection of CSI in C-RAN or F-RAN might be a more severe issue: The CSI are collected at the network edge and delivered to the cloud, more distortions are expected. Therefore, how to ensure the target Quality of Service (QoS) for different criteria, but with only inaccurate CSI knowledge, is also worth to be investigated.

Based on the research and the corresponding numerical results of this thesis, some interesting properties of C-RAN and F-RAN can be drawn, which yield some guidelines to their practical deployment in the near future.

Acknowledgement

This dissertation covers most of my research findings and achievements when I worked as a Research Assistant (Wissenschaftlicher Mitarbeiter) from 2014 to 2017 at Institut für Nachrichtentechnik of Univesität Rostock. It could not have happened without the support of a lot of people. After the completion of the last chapter, I cannot wait to express the depth of my gratitude to all of them.

First and foremost, I would like to express my sincere appreciation to my supervisor, Prof. Dr. -Ing. habil. Volker Kühn. Until today I still remember the day when I first met Volker in 2014. It was a sunny and warm afternoon in the beautiful Baltic-city Rostock, I was a little upset and nervous before meeting him for the interview, as I was quite eager to obtain his research position. However, such upset disappeared immediately just after a few words talking to him. I have never imagined that a professor can be so amiable and has the charm to warm everyone. From my first impression, he was like a old friend of me, instead of a supervisor that I first met. It was really a nice meeting and I finally obtained the precious opportunity to work in his group! From July 2014 till September 2017 I have experienced one of the most pleasant and meaningful time during my life. Besides the peaceful daily lives in Rostock and the beautiful beach of Warnemünde, I have the maximal freedom to pursue my research interest under the guidance of Volker. He never pushed me to do anything, but always encouraged me and have given uncountable valuable advises! In his research group, I never feel pressured but only energized. Moreover, I am also very grateful for the opportunities given to me to participate in numerous research activities and conferences abroad. Accompanying the call of seagulls, I have completed most research findings written in this thesis. Although I am working in Ulm nowadays, I always recall the days when I worked with Volker in Rostock. During the revision of this work, I feel again his unparalleled patience, rigor, and conscientiousness. Without him, most of my scientific achievements could not have been possible.

Secondly, I must acknowledge Prof. Dr. -Ing. Armin Dekorsy for serving as external reviewer of this dissertation. His inspiring comments, critical reviews and insightful discussions before and during the defense, have made this work more valuable. Moreover, he has invited me to give a presentation to his institutes before

the defense, I highly appreciated this chance, with which I have obtained numerous tips about how to make a good presentation. Moreover, special thanks to all the committee members of the defense chaired by Prof. Dr.-Ing. habil. Ralf Salomon.

I should also give my heartfelt thanks to the colleagues of the institute for stimulating scientific discussions and their supports. Daniel Kern, who shared the office with me, has helped me a lot not only in my research fields but also in terms of the German language. My life in Rostock could not have been so easy without the help from him. Stephan Schedler, Andre Angierski, Daniel Franz, and Henryk Richter have given me many useful advises in how to execute a good teaching task. Moreover, their own research fields have greatly broadened my knowledge. Xiang Li, Behailu Y. Shikur, Karsten Wiedmann, Nara Hahn, Prof. Dr.-Ing. habil. Tobias Weber and Prof. Dr.-Ing. Sascha Spors have also impressed me a lot with their critical thinking during each weekly internal research presentation. Stephan Lange, Kirsten Mau, Frank Jeschke and Gundula König have given me numerous supports regarding many IT and bureaucratic issues, thanks for tolerating my bad German. It is also worth to mention that I really miss the gathered lunch organized by the institute every three months, for celebrating the birthdays of the colleagues within this time period, and I also enjoy the yearly institutional BBQ in each July. Thank you all for the delicious food and the pleasant time!

In addition, I am thankful to all of my past and present teachers, tutors and schoolmates from the primary school until the university, in both China and Germany. Without your help and guidance, I cannot approach the summit of my academic career. I must mention all of them: Dezhou Tianqudonglu Primary School, Dezhou No.9 Middle School, Dezhou No.1 high School, Shandong University, Sun Yat-sen University, Universität Ulm, Ruhr-Universität Bochum, and Univesität Rostock.

Lastly my gratitude would be expressed to my beloved family for their endless love and great confidence in me all through these years. My parents, my grandparents and my wife, I love all of you forever! The written of this thesis is also accompanied by my two Ragdoll cats, thank you for being with me!

Di Chen

Ulm, September 2021

Acronyms

1G	The first Generation Mobile Network
3GPP	3rd Generation Partnership Project
4G	The 4th Generation Mobile Network
5G	The 5th Generation Mobile Network
5GPPP	5G Public-Private Partnership
ACO	Alternating Convex Optimization
A/D	Analog to Digital
AIB	Alternating Information Bottleneck
AP	Access Point
AR	Augmented Reality
BBU	Base Band Unit
BS	Base Station
CDF	Cumulative Distribution Function
CDMA	Code Division Multiple access
CF	Compute-and-Forward
CoMP	Coordinated Multi-Point
CP	Cyclic Prefix
C-RAN	Cloud Radio Access Network
CSI	Channel State Information
CSI-RS	Channel State Information Reference Signal
D2D	Device-to-Device
DCI	Downlink Control Information
DL	Downlink
EE	Energy Efficiency
eMBB	enhanced Mobile Broadband
eRRH	enhanced Remote Radio Head
EVD	EigenValue Decomposition
FBMC	FilterBank Multi-Carrier
FD	Frequency Division
FDMA	Frequency Division Multiple access
FEC	Forward Error Correction

FFT	Fast Fourier Transform
F-RAN	Fog Radio Access Network
GFDM	Generalized Frequency Division Multiplexing
HD	High Definition
IB	Information Bottleneck
IDFT	Inverse Discrete Fourier Transform
IEEE	Institute of Electrical and Electronics Engineers
IoT	Internet of Things
LTE	Long Term Evolution
MAC	Medium Access Control Layer
MCS	Modulation and Coding Scheme
MDS	Maximum Distance Separable
MF	Matched Filter
MIMO	Multiple Input Multiple Output
MINLP	Mixed Integer Non-Linear Programming
MM	Majorization Minimization
MMF	Max Min Fairness
MMSE	Minimum Mean Square Error
mMTC	Massive Machine-Type Communication
mmWave	millimeter Wave
MRC	Maximal Ratio Combining
MUX	Multiplexing
NFV	Network Function Virtualization
NNC	Noisy Network Coding
NOMA	Non-Orthogonal Multiple Access
OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal Frequency Division Multiple Access
OLM	Outer Linearization Method
OMA	Orthogonal Multiple Access
PDCCH	Physical Downlink Control Channel
PHY	Physical Layer
PUCCH	Physical Uplink Control Channel
QoE	Quality of Experience
QoS	Quality of Service

RF	Radio Frequency
RRH	Remote Radio Head
RX	Receive
SDN	Software Defined Networking
SDP	Semi Definite Programming
SDR	Semi Definite Relaxation
SE	Spectral Efficiency
SIC	Successive Inference Cancellation
SINR	Signal to Interference plus Noise Ratio
SNR	Signal to Noise Ratio
SRS	Sounding Reference Signal
TD	Time Division
TDD	Time Division Duplex
TDMA	Time Division Multiple access
TP	Throughput
TX	Transmit
UHD	Ultra High Definition
UL	Uplink
URLLC	Ultra-Reliable Low-Latency Communication
V2X	Vehicle-to-everything
VDE	Verband der Elektrotechnik, Elektronik und Informationstechnik
VR	Virtual Reality
wMMF	weighted Max Min Fairness
ZF	Zero Forcing

List of Symbols

Functions and Operators

$\lvert\cdot\rvert$	Absolute value / Cardinality
$\lvert\lvert\cdot\rvert\rvert_p$	Vector ℓ_p-norm ($p > 0$)
$\lvert\cdot\rvert_0$	Vector ℓ_0-norm
$\mathbb{E}\{\cdot\}$	Expectation of
$\succeq \mathbf{0}$	Denoting a positive Semidefinite matrix
$\mathrm{Tr}(\cdot)$	Trace of a matrix
$\mathrm{rank}(\cdot)$	Rank of a matrix
$P_{\cdot\lvert\cdot}$	Conditional probability distribution
$\Pr(\cdot)$	Probability of
$\mathrm{Diag}(\cdot)$	Diagonal matrix constructor
$\forall$	For all
$[\cdot]^T$	Transpose of a vector or matrix
$[\cdot]^H$	Hermitian transpose of a vector or matrix
$I(\cdot;\cdot)$	Mutual information
$I(\cdot;\cdot\lvert\cdot)$	Conditional mutual information
$D_{\mathrm{KL}}(\cdot\lvert\lvert\cdot)$	Kullback-Leibler divergence
$\mathrm{Var}(\cdot)$	Variance of
max	Argument of the maximum
min	Argument of the minimum
$.^{(t)}$ / $.^{(\ell)}$	t-th / ℓ-th iteration
$H(\cdot,\cdot)$	Unit step function
$\Pi_{\mathcal{C}}(\cdot)$	Euclidean projection to convex set $\mathcal{C}$
$\mathbf{1}_{L\times 1}$	A $L \times 1$ vector consisting of only element 1
$\mathbf{0}_{L\times 1}$	A $L \times 1$ zero vector

Blackboard Bold Symbols

$\mathbb{C}$	Complex number
$\mathbb{R}$	Real number

$\mathbb{N}$	Natural number

Calligraphic Symbols

$\mathcal{N}(\cdot,\cdot)$	Normal distribution
$\mathcal{N}$	Set of the eRRHs
$\mathcal{K}$	Set of the scheduled UEs
$\mathcal{X}$	Uplink transmit signal alphabet
$\mathcal{Y}$	Uplink received signal alphabet at eRRH
$\hat{\mathcal{Y}}$	Set of the compression indices
$\mathcal{M}$	Set of the requested contents in downlink
$\mathcal{G}^m$	Set of all UEs in multi-cast group requesting content f^m
$\mathcal{C}^m$	Cluster of eRRHs serving the multi-cast group $\mathcal{G}^m$
$\mathcal{C}$	Convex set
$\mathcal{P}$	Optimization problem for power minimization
$\mathcal{F}$	Optimization problem for wMMF
$\mathcal{T}$	Optimization problem for multi-cast throughput maximization
$\mathcal{R}$	Optimization problem for power allocation
$\mathcal{R}_k$	Uplink target rate in Massive MIMO based F-RAN
$\mathcal{R}_{\text{pu}}$	Uplink target rate per UE in Massive MIMO based F-RAN
$\mathcal{O}(\cdot)$	Complexity of
$\mathcal{Q}$	Number of bits required to describe CSI

Roman Symbols

B	Network bandwidth
K_{total}	Total number of UEs within the network
K	Number of the scheduled UEs
k	UE k
x_k, X_k	Uplink transmit signal from UE k
$\mathbf{x}$, $\mathbf{X}$	Uplink aggregated transmit signal vector
s_k	Uplink transmit symbol from UE k with normalized variance
$\mathbf{s}$	Uplink transmit symbol vector with normalized variance of each element
s^m	Weighting coefficient for content f^m when wMMF considered

$\mathbf{s}$	Weighting coefficient vector when wMMF considered
P_k	Transmit power of UE k
P_{UE}	Transmit power of the scheduled UE in Massive MIMO based F-RAN
P_n	Maximal allowable power of eRRH n
$\mathbf{P}$	Maximal allowable power vector among all eRRHs
p^m	Transmit power allocated for content f^m among all eRRHs
$\mathbf{p}$	Power allocation vector for requested contents
$P_{\mathrm{TX},n}$	Power consumed by the downlink transmission of eRRH n
P_{o}	Additional operational power of an active eRRH
P_{sleep}	Power consumption of a deactivated eRRH
ΔP	Difference between the operational power and the power of sleep mode
N	Number of eRRHs
n	RRH/eRRH n
L	Number of the antennas for each eRRH
l	l-th antenna
h_{nk}	Uplink channel coefficient from UE k to eRRH n
$h_{n,l}^k$	Downlink channel coefficient from l-th antennas of eRRH n to UE k
$\mathbf{h}_n^k$	Downlink channel coefficient vector from eRRH n to UE k
$\mathbf{h}_k$	Aggregated downlink channel vector to UE k
$\mathbf{H}^k$	Downlink positive Semidefinite channel matrix to UE k
$h_{n,k}^l$	Uplink small scale fading coefficient from UE k to l-th antenna of eRRH n
$\mathbf{H}_n$	Uplink small scale fading coefficients matrix from all UEs to eRRH n
$g_{n,k}^l$	Uplink channel gain from UE k to l-th antenna of eRRH n
$\mathbf{G}_n$	Uplink channel gain matrix from all UEs to eRRH n
$\mathbf{I}_{L\times L}$	Identity matrix of size $L \times L$
$\mathbf{J}_n$	Selection matrix at eRRH n
$\mathbf{J}_{n,l}$	Antenna selection matrix for l-th antenna at at eRRH n
$\widetilde{\mathbf{h}}_k$	Actual aggregated channel vector to UE k considering inaccurate CSI
Y_n	Uplink received signal at single-antenna eRRH n
$\mathbf{Y}$	Uplink aggregated received signal vector for single-antenna eRRHs
$\hat{Y}_n$	Compressed signal at single-antenna eRRH n
$\hat{\mathbf{Y}}$	Aggregated compression vector among single-antenna eRRHs
$\mathbf{y}_n$	Uplink received signal vector for at eRRH n
Z_n	Uplink additive white Gaussian noise at single-antenna eRRH n

$\mathbf{Z}$	Uplink aggregated noise vector among single-antenna eRRHs
z_n^l	Uplink additive white Gaussian noise at l-th antenna of eRRH n
$\mathbf{z}_n$	Uplink additive white Gaussian noise vector at eRRH n
$\mathbf{D}_{\mathrm{MRC},n}$	MRC detection matrix at eRRH n
$d_{n,k}$	Estimated symbol for UE k at eRRH n
$\widetilde{d}_{n,k}$	Compression of $d_{n,k}$
$\widetilde{d}_k$	Combined compressed symbol for decoding of the symbol from UE k
$C_{\mathrm{FH},n}$	Capacity of the fronthaul connecting eRRH n and the BBU pool
C_{FH}	Total available capacity when fronthauls share resources
$\mathbf{C}$	Aggregated fronthaul capacity vector
c_n	Compression rate at eRRH n
$r_{n,k}$	Required fronthaul capacity for UE k at eRRH n of Massive MIMO based F-RAN
$r_{\mathrm{FH},n}$	Required fronthaul capacity of eRRH n of Massive MIMO based F-RAN
R_k	Uplink achievable rate of UE k
t / ℓ	Iteration index
f_{LB}	Lower-Bound variable
f_{UB}	Upper-Bound variable
w_k	Weight factor for UE k
$\boldsymbol{w}$	Aggregated weight factor vector among all UEs
$\boldsymbol{g}$	Sub-gradient vector
M_{total}	Total number of the contents
M	Number of the contents being requested
f^m	Content requested by the multi-cast group $\mathcal{G}^m$
$c_n^{f^m}$	Indicator of whether content f^m is cached at eRRH n
S_n	Cache memory size of eRRH n
$v_{n,l}^m$	Beamformer coefficient at l-th antenna of eRRH n for content f^m
$\mathbf{v}_n^m$	Beamformer of eRRH n for content f^m
$\mathbf{v}^m$	Aggregated beamformer vector for content f^m
$\mathbf{V}^m$	Positive Semidefinite beamformers matrix for content f^m
$\tilde{\mathbf{v}}^m$	Normalized aggregate beamformer for content f^m among all eRRHs
$\tilde{\mathbf{v}}_n^m$	Part of the normalized aggregate beamformer $\tilde{\mathbf{v}}^m$ constructed at eRRH n
$\tilde{v}_{n,l}^m$	Part of $\tilde{\mathbf{v}}^m$ constructed at l-th antenna of eRRH n
$w_{n,l}^m$	Precoder coefficient at l-th antenna of eRRH n for content f^m
$\mathbf{w}_n^m$	Precoder of eRRH n for content f^m

$\mathbf{w}^m$	Aggregated precoder for content f^m
$\mathbf{W}^m$	Positive Semidefinite precoders matrix for content f^m
$e_{n,l}^{\text{comp}}$	Quantization noise at l-th antenna of eRRH n
$\mathbf{e}_n^{\text{comp}}$	Aggregated quantization noise vector at eRRH n
$q_{n,l}$	Standard deviation of the quantization noise at l-th antenna of eRRH n
$\mathbf{q}_n$	Standard deviation of the quantization noise vector of eRRH n
$\mathbf{q}$	Aggregated standard deviation of the quantization noise vector
$\mathbf{Q}$	Positive Semidefinite standard deviations matrix of the quantization noise
$q_{n,k}$	Quantization noise of the MRC estimated symbol from UE k at eRRH n
$Q_{n,k}$	Variance of $q_{n,k}$
$\mathbf{e}_k^{\text{CSI}}$	Aggregated CSI error vector of UE k
R^m	Downlink target rate of the content requested by $\mathcal{G}^m$
$k_{m,n}^{(t+1)}$	Re-weighted coefficient for content f^m at eRRH n in the $(t+1)$-th iteration
$u_n^{(t+1)}$	Re-weighted coefficient for eRRH n in the $(t+1)$-th iteration
u_k	Weight factor for UE k of Massive MIMO based F-RAN
$\ell_{n,k}$	Auxiliary parameter for upper-bounding the fronthaul rate of UE k at eRRH n
$T_{\boldsymbol{\beta}}$	Duration that the large scale fading coefficients stay unchanged
$T_{\boldsymbol{h}}$	Duration that the small scale fading coefficients stay unchanged

Greek Symbols

σ_n	Standard deviation of the noise at eRRH n
π	Decompression order
β	Lagrange multiplier
$\boldsymbol{\beta}$	Lagrange multiplier vector
ϵ	Tolerance factor for terminating the AIB method
ε	Tolerance factor for terminating the Bi-Section search for wMMF optimization
η	Tolerance factor for terminating sub-steps of the alternating Bi-Section method
ζ	Tolerance factor for terminating the alternating Bi-Section method
δ	Tolerance factor for terminating the Outer Linearization method
θ	Capacity allocation ratio for the uplink
ξ	Power amplifier efficiency
α	Skewness of the Zipf distribution
τ	Threshold parameter

ϵ_k	Radius of the spherical region of the bounded CSI error at UE k
δ_k	Outage probability of the bounded CSI error at UE k
Γ^m	Target SINR of the content requested by $\mathcal{G}^m$
$\boldsymbol{\Gamma}$	Target SINR vector
$\eta_{n,l}$	Auxiliary parameter for upper-bounding the soft transfer rate for eRRH n
Δ	Step size in the sub-gradient method
$\alpha_k, \beta_k, \gamma_k$	Introduced scalar auxiliary variables when the S-Lemma is adopted
$\lambda_k,\ \mu_k,\ \nu_k$	Introduced scalar auxiliary variables when the S-Lemma is adopted
ρ_{ul}	Maximum allowable transmission power in Massive MIMO based F-RAN
η_k	Power allocation factor for UE k in Massive MIMO based F-RAN
$\boldsymbol{\eta}$	Power allocation vector in Massive MIMO based F-RAN
$\mathbf{D}_{\boldsymbol{\eta}}$	Diagonal matrix constructed from $\boldsymbol{\eta}$
$\beta_{n,k}$	Uplink large scale fading coefficient between UE k and eRRH n
$\mathbf{D}_{\boldsymbol{\beta}_n}$	Uplink large scale fading coefficients matrix between all UEs and eRRH n

Publications

- **Di Chen** and Volker Kuehn, "**Alternating Information Bottleneck Optimization for Weighted Sum Rate and Resource Allocation in the Uplink of C-RAN**", *in Proceedings of 20th International ITG Workshop on Smart Antennas (WSA), Munich, Germany*, March 2016, Print ISBN: 978-3-8007-4177-9

- **Di Chen** and Volker Kuehn, "**Optimization Scheme of Noisy Network Coding in the Two Way Relay Channels**", *in Proceedings of IEEE Wireless Communications and Networking Conference (WCNC), Doha, Qatar*, April 2016, DOI: 10.1109/WCNC.2016.7565027

- **Di Chen** and Volker Kuehn, "**Scalar and Vector Compress and Forward for the Two Way Relay Channels**", *in Proceedings of IEEE 83rd Vehicular Technology Conference (VTC Spring), Nanjing, China*, May 2016, DOI: 10.1109/VTC-Spring.2016.7504460

- **Di Chen** and Volker Kuehn, "**Alternating Information Bottleneck Optimization for the Compression in the Uplink of C-RAN**", *in Proceedings of IEEE International Conference on Communications (ICC), Kuala Lumpur, Malaysia*, May 2016, DOI: 10.1109/ICC.2016.7510694

- **Di Chen**, Stephan Schedler, and Volker Kuehn, "**Backhaul Traffic Balancing and Dynamic Content-Centric Clustering for the Downlink of Fog Radio Access Network**", *in Proceedings of IEEE 17th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), Edinburgh, United Kingdom*, July 2016, DOI: 10.1109/SPAWC.2016.7536735

- **Di Chen** and Volker Kuehn, "**Weighted Max-Min Fairness Oriented Load-balancing and Clustering for Multicast Cache-Enabled F-RAN**", *in Proceedings of 9th International Symposium on Turbo Codes and Iterative Information Processing (ISTC), Brest, France*, September 2016, DOI: 10.1109/ISTC.2016.7593144

- **Di Chen** and Volker Kuehn, "**Adaptive Radio Unit Selection and Load Balancing in the Downlink of Fog Radio Access Network**", *in Proceedings of IEEE Global Communications Conference (GLOBECOM), Washington, D.C., United States*, December 2016, DOI: 10.1109/GLOCOM.2016.7841568

- **Di Chen** and Volker Kuehn, "**Joint Resource Allocation and Power Control for Maximizing the Throughput of Multicast C-RAN**", *in Proceedings of 11th International ITG Conference on Systems, Communications and Coding (SCC), Hamburg, Germany*, February 2017, Print ISBN: 978-3-8007-4362-9

- **Di Chen** and Volker Kuehn, "**An Investigation on Energy and Spectral Efficient Robust Design of Fog Radio Access Network**", *in Proceedings of 21th International ITG Workshop on Smart Antennas (WSA), Berlin, Germany*, March 2017, Print ISBN: 978-3-8007-4394-0

- **Di Chen** and Volker Kuehn, "**Robust Resource Allocation and Clustering Formulation for Multicast C-RAN with Impaired CSI**", *in Proceedings of IEEE International Conference on Communications (ICC), Paris, France*, May 2017, DOI: 10.1109/ICC.2017.7996656

- **Di Chen**, "**Low Complexity Power Control with Decentralized Fog Computing for Distributed Massive MIMO**", *in Proceedings of IEEE Wireless Communications and Networking Conference (WCNC), Barcelona, Spain*, April 2018, DOI: 10.1109/WCNC.2018.8377040

- **Di Chen**, Hussein Al-Shatri, Tobias Mahn, Anja Klein, and Volker Kuehn, "**Energy Efficient Robust F-RAN Downlink Design for Hard and Soft Fronthauling**", *in Proceedings of IEEE 87th Vehicular Technology Conference (VTC Spring), Porto, Portugal*, June 2018, DOI: 10.1109/VTCSpring.2018.8417561

Contents

Chapter 1

Introduction

This chapter contains

1.1 The Fifth-Generation (5G) Wireless System

The last few decades have witnessed an explosive growth in the wireless communications industry. The development of the cellular network from the First Generation wireless system (1G) to the 4G system is achieved not only by the innovation of RF techniques, but also with the evolution of network architecture, as well as the concepts behind it. Nowadays, the service of cellular network has been far more than just voice services, but becomes a key aspect of our daily lives with the help of Smart phones, Tablets, and Laptops, etc.. According to the investigation from Ericsson's annual report [Eri16b], the mobile data traffic has accumulated to more than 5.5 Zetabytes (5.5 billion Terabyte) per month worldwide in 2016, which has almost saturated the capacity of the current 4G network. However, lots of emerging user scenarios, such as Virtual Reality (VR), Augmented Reality (AR), Internet of Things (IoT) , Ultra High Definition (UHD) Transmission, Tactile Internet, etc.,

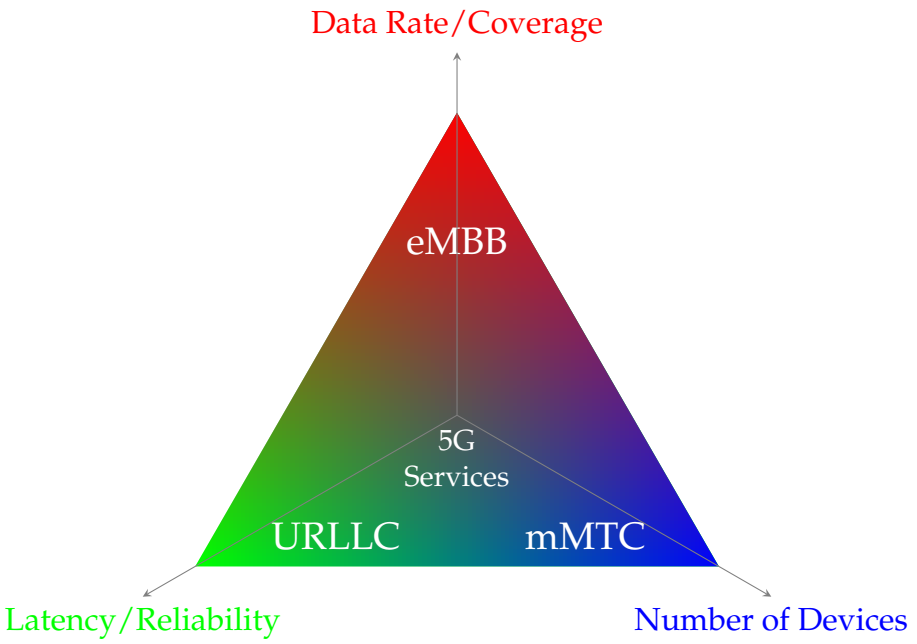


Figure 1.1: Three generic 5G services emphasizing different 5G requirements.

require even much higher data transmission rate and reliability, lower latency and energy consumption, as well as a broader network coverage.

Therefore, many global initiatives, such as 3GPP, 5GPPP, Ericsson, Nokia, Qualcomm, Samsung, etc., are collaborating on the development of 5G system and the corresponding standards. It has been agreed [OMM16] that the following three generic services should be supported by the 5G system, as shown in Fig. 1.1:

1. *Enhanced Mobile BroadBand* (**eMBB**) shall provide extremely high data transmission rates, as well as low latency for some real-time applications, e.g., VR and AR. Moreover, an extremely broad network coverage that can greatly increase users' Quality of Experience (QoE), is required to be achieved. Hence, the area capacity, which is characterized by bits/unit per area, shall be increased by roughly 1000× compared to the current LTE system.

2. *Massive Machine-Type Communication* (**mMTC**) aims to provide wireless connectivity for billions of low-cost and energy-constrained devices, so as to facilitate the concept of IoT. Therefore, the network must be able to cover immense areas seamlessly, and support the transmission for a massive number of devices. Moreover, compared with LTE, the per-link energy consumption must at least not increase. As a consequence, the target energy efficiency of 5G shall be increased by 100× at least.

3. *Ultra-Reliable Low-Latency Communication* (**URLLC**) addresses an ultra-reliable low-latency communication. More specifically, at least 99.999% service availability and reliability, with only $1-10$ ms latency [5G-15], have to be achieved

simultaneously. Such a service can bring applications such as the V2X communication, the Tactile Internet, into reality.

In order to fulfill the above challenging requirements, i.e., increasing both Energy Efficiency (EE) and Spectral Efficiency (SE) simultaneously with higher reliability and reduced latency, both the RF techniques and the network architecture have be evolved and even revolutionized. In Physical Layer, new waveforms such as GFDM [FKB09] and FBMC [MBe10] have been discussed, so as to overcome some limitations and drawbacks of OFDM. A straightforward approach to increase the network capacity is to increase the bandwidth used for transmission, hence, the mmWave frequency bands [Ne15] ranging from 30 GHz to 300 GHz are under investigation recently. Another straightforward approach is simply increasing the number of antennas: It has been shown in [Mar10; Rus+13; Mar+16; NCS17] , with a massive number of antennas, and by exploiting the resultant *channel hardening* effect, the resultant Massive MIMO is scalable and can lead to huge performance improvement, in terms of both SE and EE, without incurring too high complexity and too much amount of overhead. Moreover, the Full-Duplex Communication [Son+17], with which the signals are transmitted and received in the same frequency band simultaneously, can theoretically double the current SE immediately, compared with the conventional Half-Duplex mode. Furthermore, from 1G to 4G, only Orthogonal Multiple Access (OMA) is adopted, i.e., FDMA, TDMA, CDMA, or OFDMA. While from the perceptive of the information theory, for given amount of transmission resources, e.g., time or frequency, the Non-orthogonal Multiple Access (NOMA) always outperforms the OMA [GK11], due to its more efficient usage of the available resources. Hence, NOMA is also discussed for 5G [Dai+15].

Besides the above-mentioned innovative techniques, rethinking of the network architecture is also a promising direction, which can forecast even more performance improvement, as well as lower cost. Therefore, the concepts of the Network Function Virtualization (NFV) [AT12], and the Software Defined Networking (SDN) [Fou12], with which much more flexibility and scalability in future networks can be achieved, are to be utilized. In particular, the Cloud Radio Access Network (C-RAN) [Mob11], as well as the Fog Radio Access Network (F-RAN) [Pen+16] have been shown to be promising architectures and platforms to run NFV and SDN [Won+17]. In C-RAN, the radio connectivity to mobile users is provided via densely deployed low-cost Remote Radio Heads (RRH), where only basic RF functions are executed. The RRHs act only as RF signal collectors and emitters: In the uplink, they forward the collected signals to the Base Band Units (BBU) pool in the cloud, via the fronthauls. In the downlink, they receive the pre-processed signals and emit them without further processing. The servers located in the cloud with strong computational capabilities undertake most of the base-band signal processing functionalities in a centralized manner. With such a centralized joint signal processing,

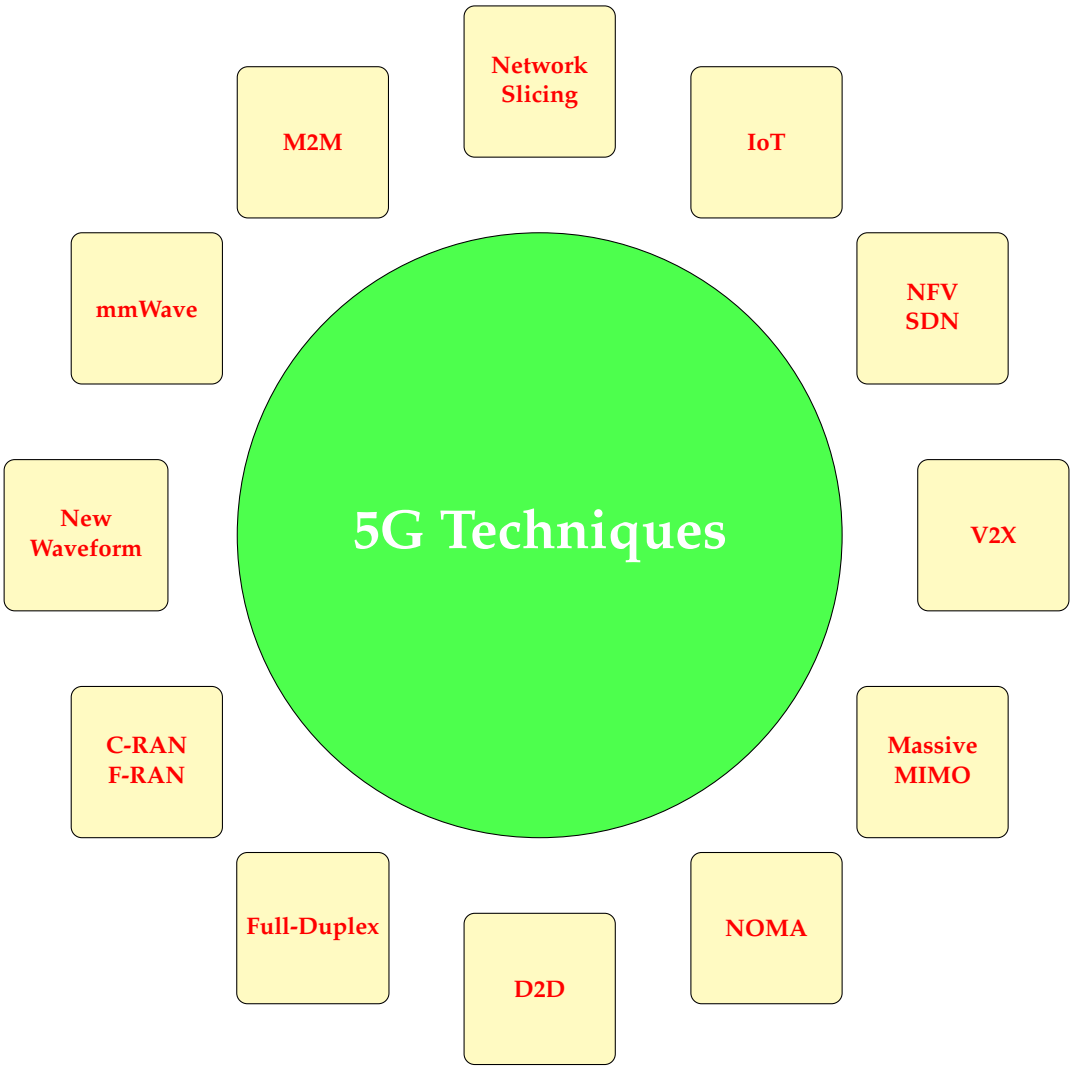


Figure 1.2: An illustration of 5G techniques and concepts.

much more efficient interference management, transmission coordination, load balancing, resource allocation, etc., can be achieved [Pen+15; Que+17], which are able to significantly increase both SE and EE of the network. Moreover, the centralization can better coordinate the inter-user interference arising from NOMA, and as stated above, provide an ideal platform for running various Virtual Network Functions of NFV and different Layers of SDN [OMM16]. However, the fully centralized processing might incur extremely high computational complexity, a large amount of overhead, and intolerable latency. Hence, F-RAN is proposed, where the edge computing is introduced at the network edge, e.g., RRHs, with which partial functionalities can be undertaken there instead of the cloud. Hence, the heavy burden on the fronthauls and cloud servers can be relieved [Wue+14]. Naturally, such a functional split leads to a trade-off between the computational complexity and the performance improvement. Intensive illustrations and discussion will be given in the main part of the thesis.

Data Offloading, Unlicensed LTE, D2D transmission, etc., are among the other promising concepts and techniques for 5G [Won+17]. Due to space limitations we can not elaborate on all of them, an illustration of several 5G techniques and concepts can be seen in Fig. 1.2.

1.2 Cloud Radio Access Network (C-RAN)

One of the main focus of this thesis is C-RAN. C-RAN was firstly proposed by China Mobile [Mob11] in 2011 and quickly draws the attention from the researchers worldwide [Par+13b; Par+14; SZL14; Wue+14; ZY14; Pen+15; Tao+16; Que+17]. By incorporating the concept of the cloud computing into the traditional Radio Access Network (RAN), it proves to be the most promising network architecture to meet the challenging demands of 5G. In C-RAN, a traditional Base Station (BS), as well as the functionalities it undertakes, is decoupled into two parts: the **Remote Radio Head (RRH)** and the **Base Band Units (BBU) pool**, these two parts are connected via the fronthaul.

- **Remote Radio Head**: The RRHs are low-cost Access Points (APs) for the User Equipment (UE). They are densely and ubiquitous deployed within the network. These *stupid* APs perform only basic RF functions, such as the Analog-to-Digital conversion, the Digital-to-Analog conversion, etc.. Hence, they can be deployed in a large scale but without incurring too much costs. Compared with LTE, a large number of RRHs can provide a seamless network coverage and greatly shorten the distances between the UEs and APs. Such a short distance is a straightforward and most effective approach to increase the per-link

SE. Moreover, due to the attenuation properties of the mmWave [Ne15], the densely distributed RRHs can also facilitate the realization of the mmWave communication. Together with the massive number of low-cost APs, such a deployment can greatly increase the area capacity to meet the demand of 5G targets.

- **The Base Band Units pool**: It is remotely located in the cloud and in charge of all RRHs. The centralization enables joint signal processing, coordinated interference management, optimized network resource allocation and scheduling, etc.. From the viewpoint of the BBU pool, the RAN is actually a large-scale virtual Multiple Input Multiple Output (MIMO) system. Hence, a Networked Coordinated Multi-Point (CoMP) transmission can be easily realized.

- **Fronthaul**: A fronthaul connects a specific RRH and the BBU pool. It can forward the RF signal from the RRH to the cloud and vice versa. The fronthaul can be constructed via different technologies, such as the optical fiber communication (wired fronthauling), or the millimeter wave communication (wireless fronthauling) [Pen+15]. The optical fiber connection provides high capacity at the expense of higher cost and inflexible deployment of RRHs. Compared with the optical fiber, the wireless fronthauling has lower capacity, less reliability, and the resources have to be shared among RRHs, but it is much cheaper and can facilitate a flexible deployment. According to [DC15], for a dense or heterogeneous network, the wired fronthaul is usually not feasible, its wireless counterpart is the practical solution in such scenarios.

An illustration of C-RAN is shown in Fig. 1.3, where its connectivity to the core network is also depicted.

1.2.1 Uplink

The uplink transmission of C-RAN denotes the delivery procedure of the information from the scheduled UEs, via RRHs and fronthauls, to the BBU pool in the cloud. The whole procedure consists of two hops, i.e., the **Radio Access Hop** and the **Fronthauling Hop**, and two processing sites, i.e., the **RRH Processing (edge)** and the **BBU pool Processing (cloud)**.

1. **Radio Access Hop**: In this hop, the scheduled UEs encode and modulate their independent information into analog signals, and send them. Such information are intended for the cloud to decode. The radio resources, e.g., the time and frequency resource, are shared among all UEs. Therefore, the UEs interfere with each other, and all RRHs receive different superposition of the signals from all scheduled UEs.

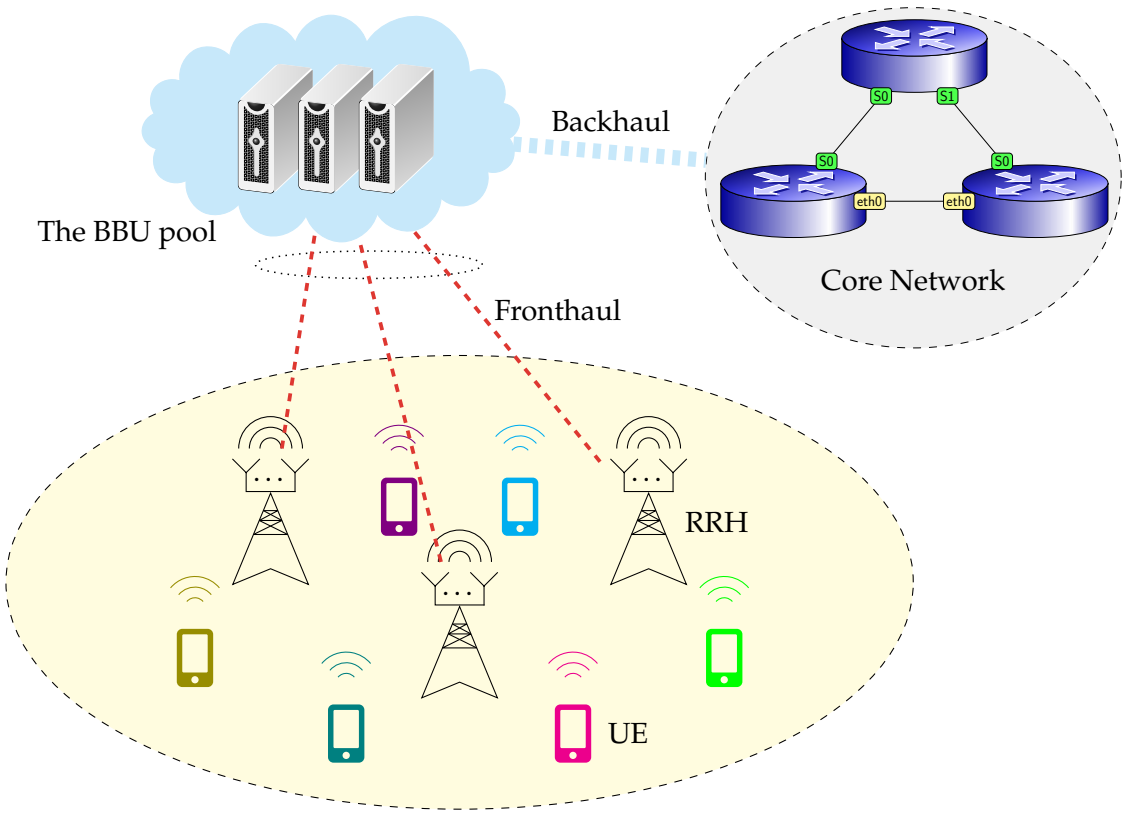


Figure 1.3: An illustration of the Cloud Radio Access Network.

2. **RRH Processing (edge)**: In the ideal case, the received superposed analog signals should be delivered by fronthauls in the next hop to the cloud, without any further processing at RRHs. Obviously, the delivery of the analog signals without any distortion requires the fronthaul with infinite capacity. Thus, sampling and Analog-to-Digital (A/D) conversion at RRHs are inevitable. According to [Par+14], even with the current LTE configurations, when a RRH with two antennas serves three cell sectors using five carriers, and the A/D converter adopts a standard scalar quantization technique with 15 bits/baseband IQ sample, the capacity of the fronthaul link must at least 10 Gbit/s! With the network configuration of 5G , such a value can be even much higher, which is infeasible for low-cost and densely distributed RRHs. Therefore, in addition to the A/D conversion, further compression of the digital signals at RRHs is necessary. The compression procedure should be optimized to exploit the available capacity of its connecting fronthaul, and retain as much useful information at the destination as possible.

3. **Fronthauling Hop**: In this hop, the compressed signals are delivered via the corresponding fronthauls to the cloud. For the wired fronthauls, e.g., the optical fibers, these signals have their own fronthauling resources. However, for the wireless fronthauls, e.g., the mmWave, the fronthauling resources have to be shared among all RRHs. Thus in this scenario, the optimization of the re-

source allocation shall also be taken into consideration. It should be noted that the fronthaul resource allocation will also influence the optimization of compression, i.e., the optimization of the compression process and the resource allocation interact with each other. Hence, a joint consideration of them is required. As we are going to show later, this is also one of the main contributions of this dissertation.

4. **BBU pool Processing (cloud)**: At the BBU pool in the cloud, the received compressed signals from all RRHs are decompressed firstly. Note that due to the independent superposition of all UE signals at all RRHs, each compressed signal received by the cloud contains certain information from each UE. Hence, for a better information retrieval, the information from the same UE shall be combined before decoding. In order to retrieve the information for each UE from the combined signal, a specific detection step, e.g., Matched Filter (MF), Zero Forcing (ZF), or Minimum Mean Square Error (MMSE), is to be performed. After the signal detection procedure, the decoding of the original information is then followed. From the perspective of information theory, a joint decompression, detection and decoding is optimal, which will, however, definitely result in much higher complexity. More details will be given in the coming chapters.

It is worth to mention that the global CSI should be accessible at the BBU pool in the cloud, so as to obtain an optimal compression strategy, fronthaul resource allocation, joint decompression and detection of signals. Therefore, a large amount of overhead is inevitable, which is also a key difficult for the practical realization of the C-RAN. We will address this issue later in detail.

1.2.2 Downlink

The downlink transmission of C-RAN features the delivery procedure of the information from the BBU pool in the cloud, via the fronthauls and RRHs, to the scheduled UEs. Similar to the uplink, the whole procedure also consists of two hops, and two processing sites. However, the signal processing tasks undertaken by each part are far more different from that in the uplink.

1. **BBU pool Processing (cloud)**: From the viewpoint of the cloud, all RRHs form a virtual networked MIMO system. Hence, the cloud can process the information intended for each UE as if a real MIMO system exists, e.g., the power control, beamforming, etc., can be considered in a similar way. Moreover, the signal construction procedure should also take the capacity-limited fronthauls into consideration.

2. **Fronthauling Hop**: In the downlink, according to how signals are processed by the BBU pool, two modes of fronthauling strategies are adopted, i.e., the *soft* transfer mode and the *hard* transfer mode. The soft transfer mode represents a compression-based strategy [DY16b]. Here, the BBU pool forms the complete base-band signals to be transmitted by the RRHs. It includes the encoding and the modulation of the requested data, as well as the RRH specific spatial precoding. These signals are superposed, compressed and transmitted to RRHs via fronthauls. Obviously, such a signal compression step is required to be optimized. Contrarily, the hard transfer mode refers to a data-sharing strategy [DY16b]. Here, raw encoded data streams are separately forwarded via fronthauls to different subsets of RRHs. This is due to the fact that, it might be impossible to forward all data streams to all RRHs via capacity-limited fronthauls. Hence, the cluster pattern, which describes which subset of RRHs (cluster) should serve which UE, is subject to be optimized. The downlink signal compression in the soft transfer mode and the cluster formulation strategy in the hard transfer mode will be intensively addressed in later chapters.

3. **RRH Processing (edge)**: When the soft transfer mode is adopted, the RRHs decompress the received signals and simply forward them to UEs, without any further processing, as they have been already modulated and precoded in the cloud. While with the hard transfer mode, the RRHs should decode the received raw data streams, then beamform, modulate, and send them.

4. **Radio Access Hop**: In this hop, the signals are transmitted by RRHs and received by the scheduled UEs.

Similar to the uplink, the global CSI is also required at the BBU pool in the cloud, for the signal processing and the network optimization.

1.2.3 State of the Art

For the uplink of C-RAN, most works focus on how to design quantizers at RRHs for the compression step. In [Par+14], the performance of the point-to-point compression, distributed compression exploiting the Wyner-Ziv coding [WZ76], and Compute-and-Forward (CF) are compared. It shows that the performance advantage of the distributed compression over the point-to-point compression increases as the Signal to Noise Ratio (SNR) becomes higher. Moreover, CF can outperform all the other schemes, as the SNR falls into the regime where the fronthaul capacity becomes the main performance bottleneck. In [ZY14], a new optimization mechanism for the Wyner-Ziv coding based compression is proposed, showing that by setting the quantization noise levels to be proportional to the background noise levels, the

compression steps can approach optimality. An OFDMA-based C-RAN system is consider in [LBZ15], where a practical uniform scalar quantization mechanism in the uplink is proposed. When it comes to multi-hop fronthauls, routing and in-network processing schemes are discussed in [Par+14], and several compression strategies are proposed and compared in [Par+16]. Considering the processing step at the BBU pool, a joint decompression and decoding strategy is investigated in [Par+13b].

For the downlink of C-RAN, the published works can be coarsely classified according to the adopted fronthauling modes. When the hard transfer mode is considered, the construction of beamformers is investigated in [SZL14], in which the algorithm to optimize beamformers for energy efficient downlink C-RAN is proposed. The issue of user-centric RRH clustering is discussed in [DY14]. For a given fixed cluster pattern under per-RRH power constraints, the beamformer construction is considered in [DY15]. When it comes to the soft transfer mode, different compression optimization schemes are proposed in [Par+13a; DY16b]. The performance comparison of these two modes can be found in [PDY15; DY16b]. When the fronthaul network is multi-hop and has certain topology, the fronthauling scheme and the network optimization are discussed in [AS16; LY17], where the beamformer construction and a network coding based fronthauling are proposed respectively. The issue of the Signal to Interference plus Noise Ratio (SINR) balancing in the downlink is investigated in [LZ16].

1.3 Fog Radio Access Network (F-RAN)

Compared with the current wireless network architecture, although lots of benefits provided by the C-RAN have been demonstrated [Pen+15; Que+17], some limitations and disadvantages are also followed. One of the most significant issue of C-RAN is its high demand on the fronthauls. This issue arises mainly from the fact that, C-RAN pushes almost all base-band signal processing functionalities to the BBU pool in the cloud. Although it can be partly overcome by the compression step at RRHs, sometimes ultra-high capacity might still be required, in order to guarantee certain level of Quality of Service (QoS) and QoE. As we have stated before, one key feature of 5G network is the ultra densely deployed low-cost APs in order to greatly increase the network coverage and decrease the distance between UEs and APs. Thus, such a high demand on fronthauls would also result in difficulties on such a deployment. Another problem of the C-RAN, is the requirement of the global CSI knowledge at the BBU pool in the cloud, so as to make it possible, to perform almost all steps of base-band signal processing in a coordinated manner, as well as to design and optimize the whole network. Hence, a large amount of overhead is

inevitable, and sometimes it might even counterbalance the benefits of the C-RAN completely. Furthermore, the fully centralization puts also much computational burden on the cloud server, and can incur unacceptable latency in delay-sensitive services.

To overcome several disadvantages of the C-RAN, the Fog Radio Access Network (F-RAN), exploiting the fog computing, or in another word, edge computing, has been recently proposed and widely discussed [Bon+14]. Contrary to the cloud computing, the fog computing enables certain functionalities still to be executed at the network edge, e.g. APs or even UEs, instead of only at the remote servers. In particular, a substantial amount of storage, communication, control, configuration, measurement and management are *pulled back* from the cloud to the network edge again [Pen+16], the RRH becomes the so-called **enhanced RRH (eRRH)**. Therefore, the fog computing reduces the distance between the computing modules and UEs, and this why **Fog** is used to name such an architecture. F-RAN can be regarded as a combination or a compromise between the traditional network architecture and the C-RAN. It can avoid several difficulties in the practical deployment of the C-RAN, e.g., high burden on fronthauls and the cloud server, and retain several key features and advantages of it, as the partial centralization is still kept. However, on the other hand, a theoretical performance loss compared to the C-RAN is thus inevitable. Hence, the trade-off between the network performance, and the hardware requirements as well as the computational complexity should be taken into consideration when such networks are designed. Which and how many functionalities can be pulled back and implemented at the network edge, are tightly dependent on the service requirements, the hardware conditions, etc..

According to the descriptions above, the F-RAN can be constructed based on Fig. 1.3, as shown in Fig. 1.4. We see that the RRHs are equipped with either a cache module or a processor. Actually these are two approaches that are widely discussed to realize the fog computing.

1.3.1 Caching

Recent studies [Pou+16; Ara+17] show that popular multimedia streams with high data rate requirement, e.g., the newly released HD movies, live sport matches, etc., would generate a significant portion of the whole network traffic. Moreover, this is a typical user scenario in the future 5G system. The same contents might be requested by many users simultaneously. Hence, introducing a cache module on edge devices but retaining all other base-band processing functionalities still at the BBU pool in the cloud is a cheap and easy, but an effective way for a specific realization of F-RAN, as shown in Fig. 1.4: Some RRHs are equipped with a **Cache Module**, and

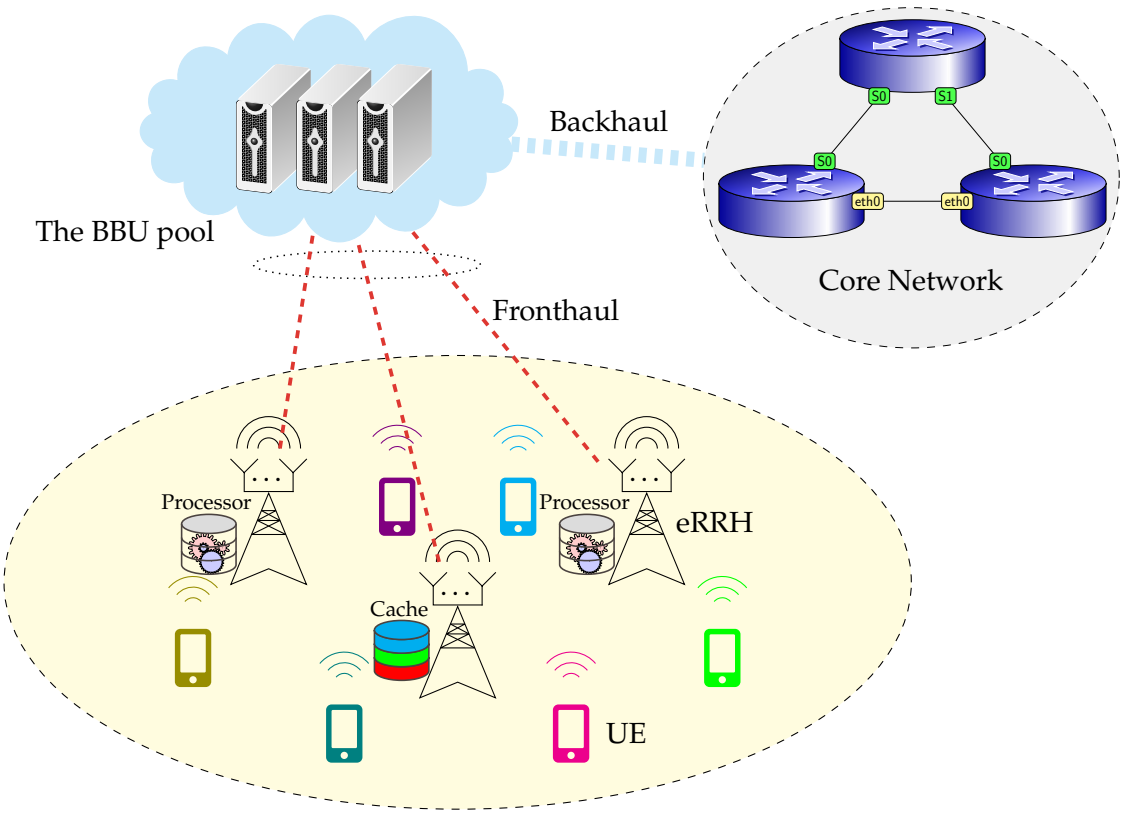


Figure 1.4: An illustration of the Fog Radio Access Network evolved from the Cloud Radio Access Network depicted in Fig. 1.3.

the other RRHs are equipped with a **Processor**, which can undertake certain amount of base-band signal processing functionalities to realize the fog computing. In both cases, RRHs evolve into eRRHs. Specifically in the first case, by caching some popularly requested contents at eRRHs at the *off-peak* time, the downlink transmission of these contents would not consume the fronthaul resources anymore. As a consequence, the traffic burden on fronthauls at the *peak* time can be greatly reduced [Sha+13; Wan+14]. Moreover, the unequal popularity and the multi-cast nature, i.e., some contents can be rather probable to be requested by lots of UEs, make caching some popular contents more reasonable. In addition to reducing the burden on fronthauls, caching can also reduce the outage probability of QoS, and improve the robustness of the network. More details will be given later.

In order to achieve an effective cache placement, M. A. Maddah-Ali and U. Niesen's pioneering work [MN14] provides the upper and lower bounds of the capacity of the caching system, from the perspective of the information theory. It theoretically confirms that the network capacity can be improved further with the help of caching. In their work, two schemes are proposed, i.e., the uncoded caching and the coded caching. With the uncoded caching, complete files are cached. While with the coded caching, different fractions (e.g. parity bits) of the files are stored at different cache modules using the Maximum Distance Separable (MDS) codes, e.g. Fountain code. Furthermore, D. Gündüz etc. propose a proactive content caching strategy

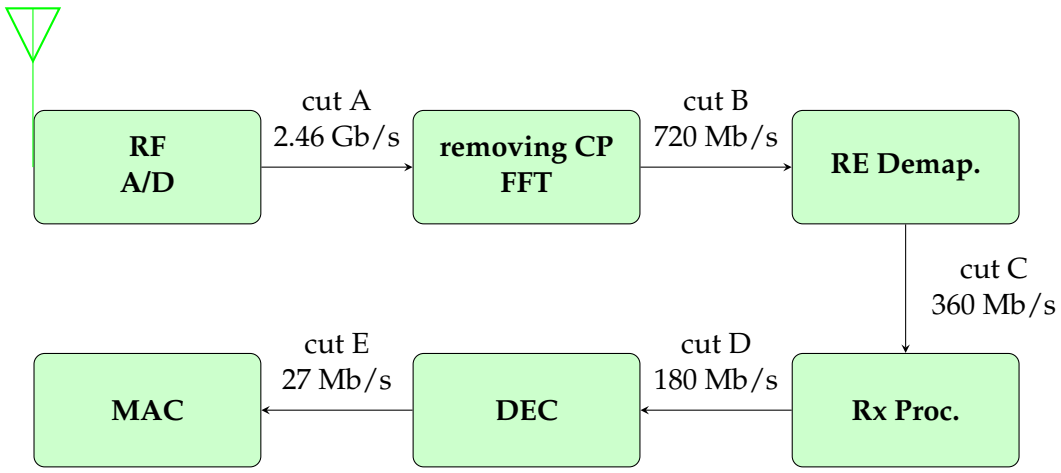


Figure 1.5: The different functional splits and the corresponding required fronthaul capacities: **RE Demap.**: Resource Element De-mapping; **Rx Proc.**: Receive Processing (incl. frequency domain equalization, Inverse Discrete Fourier Transform (IDFT), etc.); **DEC**: Forward Error Correction (FEC) decoding; **MAC**: Medium Access Control Layer [Wue+14].

that can even outperform the reactive caching strategy [SGG18].

1.3.2 Flexible Functional Split

Besides caching contents at the network edge, a more general way to implement the fog computing is to pull some functionalities back to the network edge again. Based on the results from [Wue+14], in an OFDM-based C-RAN, when only basic RF and A/D functions are performed at RRHs, the I/Q symbols including the Cyclic Prefix (CP) should be transmitted by the fronthaul to the cloud. As almost no signal processing procedures are executed at RRHs, they can be potentially constructed in very small sizes and the costs can be quite low. This is equivalent to splitting the whole signal processing chain at cut *A* in Fig. 1.5. According to the system configuration and the corresponding computation described in [Wue+14], when the function is split at this point, the required fronthaul capacity is at least 2.46 Gbit/s per fronthaul link. When a RRH evolves into an eRRH, by undertaking the task of removing CP and doing FFT, i.e., the function is split at cut *B*, the required fronthaul capacity can be then reduced to 720 Mbit/s. Similarly, if more and more functionalities are executed by the eRRHs, the required fronthaul capacity can be further reduced, but the network becomes more and more close to the traditional network architecture, and the performance benefits arising from the centralization will diminish. Hence, facing different demands of the 5G services in future, as well as the variation of the network conditions, a flexible PHY (Physical Layer) functional split in F-RAN is a promising technique to deal with these issues. Furthermore, it is also an enabler to run NFV and SDN [OMM16].

1.3.3 Uplink

Similar to the C-RAN, the uplink transmission of the F-RAN also consists of two transmission hops and two processing sites.

1. **Radio Access Hop**: This procedure is similar to the C-RAN.
2. **eRRH Processing (edge)**: Based on the configuration of the functional split, eRRHs undertake the corresponding signal processing tasks. After this procedure, the output data might be further compressed in order to accommodate with the capacity of the fronthaul. It should be noted that the more functionalities are undertaken by eRRHs, the higher costs of eRRHs are expected, but the costs of the fronthauls can be reduced, as less capacity is required.
3. **Fronthauling Hop**: After the signals are processed by eRRHs, the fronthauling of these signals is similar to that of the C-RAN,
4. **BBU pool Processing (cloud)**: The BBU pool decompresses the received signals jointly or separately at first, then it performs the rest functionalities that are not executed at the eRRHs.

1.3.4 Downlink

Compared with the C-RAN, the downlink transmission of the F-RAN consists of the same two transmission hops and two processing sites.

1. **BBU pool Processing (cloud)**: Based on the configuration of the functional split, the BBU pool undertakes the corresponding signal processing tasks. Note that if the eRRHs are equipped with cache modules, and several requested contents have been cached, the BBU pool do not need to construct and process the signals for these contents. After the signal construction and the processing steps, the output data should be further compressed in order to accommodate to the capacity of the fronthaul.
2. **Fronthauling Hop**: After the signals are processed by the BBU pool, the fronthauling of these signals is similar to that of the C-RAN,
3. **eRRH Processing (edge)**: After the eRRHs receive the signals from the fronthaul, they perform decompression to reconstruct the signals. Then all remaining functionalities that are not carried out by the BBU pool would be performed on these signals. For all requested contents that are cached locally,

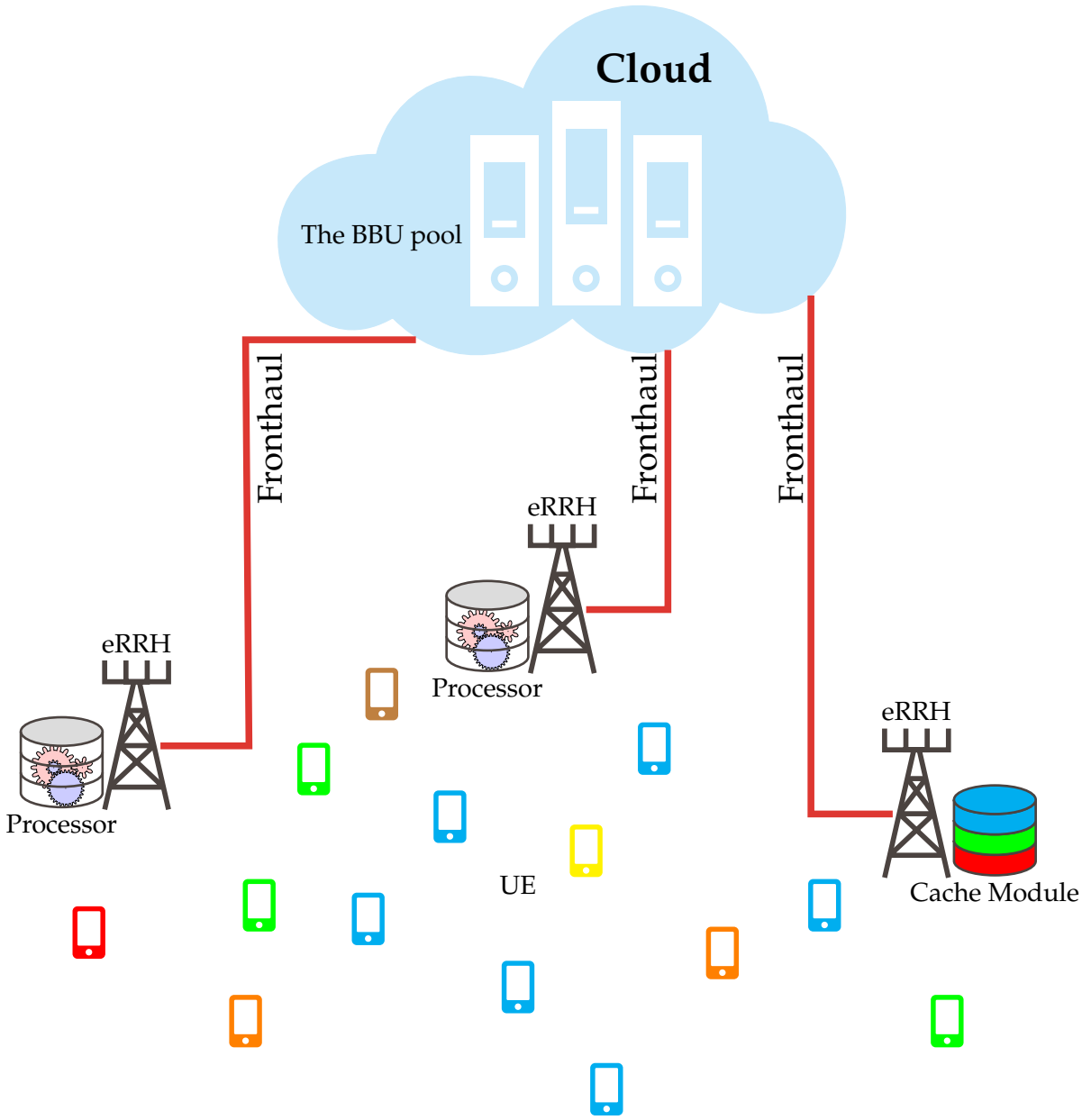


Figure 1.6: An illustration of the cache-enabled F-RAN under the multi-cast scenario. The UEs with the same color denote that they request the same content.

the signal construction and the whole processing chain are to be done at eRRHs. It should be mentioned that as the cached contents are processed locally at eRRHs, no signal distortion occurs compared to the other information that are compressed and transmitted via the fronthauls.

4. **Radio Access Hop**: This procedure is similar to the C-RAN.

As we only investigate the F-RAN part without its connection to the core network, we mainly focus on the left part of Fig. 1.4. Hence, we adopt Fig. 1.6 as the base model for future investigation in this thesis.

1.3.5 State of the Art

Since caching can greatly reduce the computational burden on the BBU pool and the transmission burden on fronthauls, and it is a simple and low-cost way to achieve F-RAN, lots of work focus on the design of the cache-enabled F-RAN. In [Pen+14], a joint design of the cache content placement and downlink beamformer is investigated, aiming to minimize the network energy cost including both eRRHs and the fronthaul. A cooperative transmission and caching scheme are investigated in [Che+16]. From the perspective of the information theory, a proactive caching scheme is proposed in [Gre+15]. For the multi-cast scenario, when the uncoded caching scheme is adopted at eRRHs, an efficient high EE oriented networked beamformer construction algorithm is proposed in [Tao+16]. For the coded caching scheme, a similar algorithm is shown in [UAS16]. Furthermore, a joint optimization of the cloud and fog processing procedures for F-RAN is summarized in [PSS16].

As for the functional split, different splitting options with the corresponding fronthaul requirements are computed and summarized in [Wue+14]. The performance comparison of different splits can be found in [DLG16]. From the viewpoint of the industry, the feasibility of both PHY and MAC layer functional split is investigated in [Mou+17]. More intensive study for the functional split in 5G gNB can be found in [Eri16a].

1.4 Massive MIMO

Another key technology for the 5G networks is Massive MIMO [Mar10; Mar+16], where the number of antennas equipped on the BS is significantly larger than the number of the served users or data streams, as shown in Fig. 1.7. It has been demonstrated that a network operating in the regime of Massive MIMO has several

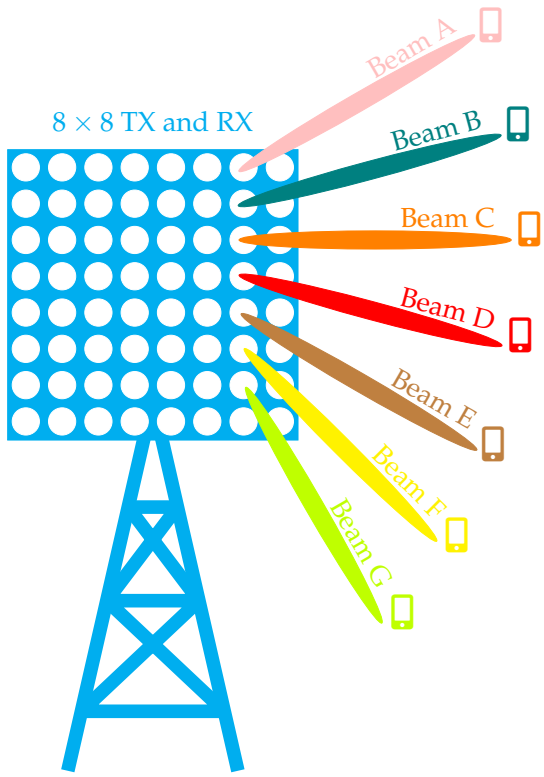


Figure 1.7: A Base Station equipped with a 64-antenna Massive MIMO.

advantages [Mar+16]: Firstly, both SE and EE of the network can be significantly increased, this is due to the fact that with so many antennas, the beams can be generated significantly narrow and more directed to each user. Hence, the interference between different data streams can be greatly reduced, the energy consumption can thus be decreased. Secondly, compared with the traditional multi-user MIMO, where the CSI is required at both sides of the BS and the users, in the Time Division Duplex (TDD) Massive MIMO, by exploiting the reciprocity of the channel, the CSI is not necessary to be measured by users anymore. Such a property can significantly reduce the amount of downlink pilot signals transmitted by the BS to the users. Hence, a Massive MIMO system is scalable, as the number of pilot signals relies only on the number of users, instead of the number of antennas [Mar+16]. Thirdly, when the number of antennas is sufficiently large, an effect known as *channel hardening* takes place, due to the law of large numbers. Under such a situation, the effects of the small-scale fading and the frequency dependence will disappear. Then from the perspective of a user, the radio link between itself and the BS becomes rather close to a deterministic scalar channel, with known, frequency-independent channel gain and additive noise [Mar+16]. Therefore, the signal processing procedure, resource allocation, user scheduling, etc. can be greatly simplified. More detailed introduction and demonstration of such advantages can be found in [Mar10; Mar+16].

However, one problem of Massive MIMO is the performance degradation when the number of antennas decreases. If not so many antennas can be mounted on a BS, the system becomes more and more close to a traditional multi-user MIMO, and thus loses the properties and advantages of Massive MIMO. On the other hand, by increasing the number of the equipped antennas, a more powerful Massive MIMO system can thus be realized. Nevertheless, the size of a BS usually limits the maximum number of its antennas.

1.5 Networked Massive MIMO based F-RAN

As introduced above, both C-RAN/F-RAN and Massive MIMO have their advantages and limitations. As for the C-RAN, although high SE and EE feature this system, and the low-cost RRH can be easily deployed densely, the fully centralized signal processing, scheduling and optimization would impose heavy computational burden on the BBU pool, and extremely high capacity of the fronthaul is required. Moreover, the request of the global CSI in the cloud leads to lots of overhead and high latency. These issues become more severe when more RRHs exist in the network. Although F-RAN can partially relieve such a burden, the global CSI is still required at the BBU pool to perform the network design and optimization. When the number of eRRHs become larger, the introduced overhead might still overwhelm the benefit of F-RAN [Par+14; Pen+16; PSS16; Tao+16]. Hence, a practical implementation approach for C-RAN and F-RAN, with which their theoretical benefits can be kept and realized, is urgently needed.

Massive MIMO also features high SE and EE, as well as the simplified signal processing procedure, scheduling, etc.. Moreover, the amount of overhead for the CSI can be greatly reduced, as the influence of the small-scale fading disappears due to the effect of channel hardening. However, as introduced above, the existence of such advantages is closely dependent on the number of equipped antennas. When less antennas are mounted, the benefits of Massive MIMO vanish rather rapidly. Unfortunately, 5G network features a dense and low-cost deployment of BSs, which might contradict with the requirements of Massive MIMO.

In order to overcome the disadvantages and difficulties of these two techniques, and even to boost their advantages to each other as well, we consider a combination of them, as shown in Fig. 1.8. We call such a system a Networked Massive MIMO based F-RAN, whose architecture is similar to F-RAN. However, each eRRH is equipped with more antennas, but this number can be smaller than a single Massive MIMO system. Similar to the F-RAN, each eRRH has limited computational capabilities to perform the fog computing. By exploiting the benefits of F-RAN,

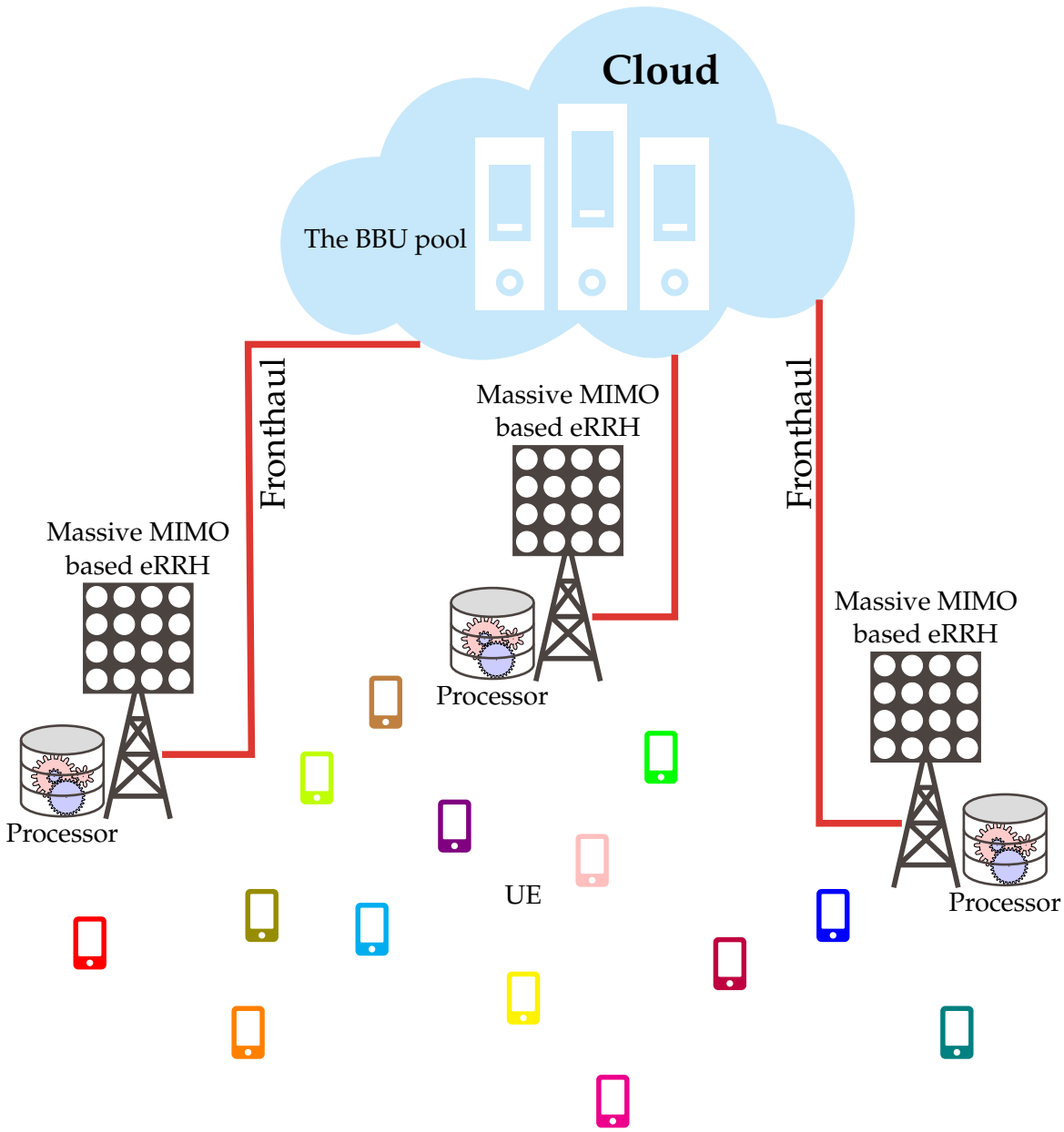


Figure 1.8: A Networked Massive MIMO based F-RAN.

from the perspective of the BBU pool, all eRRHs actually form a networked Massive MIMO system, or a distributed Massive MIMO [SYC14; PCB15]. Hence, these two techniques might benefit from each other and overcome their own shortcomings. For example, too many antennas are not necessarily to be mounted on a single eRRH, and as we are going to show later, we extend the works [SYC14; PCB15], by proposing a low complexity and partially distributed network optimization and signal processing mechanism, with which the amount of overhead and the computational burden on the BBU pool can be greatly reduced.

1.6 Outlines and Contributions

In Chapter 2, we are going to introduce some preliminary information and mathematical tools, which will be utilized later: The rate distortion theory and the Information Bottleneck (IB) method are firstly introduced. Then we prepare some optimization tools and techniques for designing and optimizing the network for future use.

In Chapter 3, we investigate the network design for the uplink of C-RAN and F-RAN. As introduced before, the high capacity requirement on the fronthaul is the key limitation from putting C-RAN into practical use. Although F-RAN can lower the traffic on fronthauls by exploiting the fog computing, compressing the signals received by eRRHs is always beneficial for reducing the demand on it. Hence, the quantizers used for realizing the compression play an important role in the alleviation of the fronthaul burden. As there are multiple eRRHs receiving correlated signals in C-RAN/F-RAN, we extend the well-known IB method, which is used for the case of single-quantizer, to a so-called Alternating Information Bottleneck (AIB) method, with which a new algorithm for joint optimizing the compression steps executed at RRHs/eRRHs is proposed. Moreover, in case the fronthaul resources have to be shared and dynamically allocated among RRHs/eRRHs, the AIB method can also be adopted, for the optimization of the resource allocation on the fronthaul. We also analyze the convergence behavior of the proposed algorithm, and provide numerical results to demonstrate the effectiveness and correctness of it.

In Chapter 4, we consider the network optimization for the downlink of C-RAN and F-RAN. As stated in Subsection 1.2.2, there are mainly two different data sharing strategies in the downlink of fronthaul transmission, i.e., the hard and the soft transfer mode. For the hard transfer mode, it is essential to optimize the cluster formulation of RRHs/eRRHs for serving different UEs in the uni-cast scenario, or for serving different groups of UEs in the multi-cast scenario. At the same time, the resultant downlink traffic on each fronthaul must be supported. We propose an

optimization algorithm, where the cluster formulation and the traffic balancing are simultaneously taken into account. For the soft transfer mode, the key procedure is the compression, and the precoder design. Again, a joint optimization mechanism for the compression and the precoder generation is proposed. Furthermore, both high EE and SE oriented network work design are considered in our work, in order to accommodate to different service requirements. For high EE oriented design, we consider not only the transmission power, but also all additional operation power of an active RRH/eRRH. Therefore, it is shown that in some cases, switching off some RRHs/eRRHs might save more power, even at the price of more transmission power consumption. The results can be a meaningful operational guideline for the network provider. For high SE oriented design, joint power allocation and beamformer construction approaches are investigated for different criteria, i.e., the Throughput Maximization, and Max-Min Fairness. Additionally, the robust design is also to be studied when only inaccurate CSI is available at the BBU pool. As we are going to show later, the propose robust design mechanism can work for both hard and soft transfer mode, and certain QoS can always be guaranteed even only inaccurate CSI is present. In the end, some numerical results are provided based on the proposed algorithms.

Up to now, the network design and the optimization are centrally executed by the BBU pool for both C-RAN and F-RAN. Hence, the global CSI is required, which can incur lots of overhead and greatly reduce the system capacity in practice. Moreover, the complexity of the centralized design is rather high. Therefore, in Chapter 5, we try to tackle these issues by introducing a combination of the concept from Massive MIMO, and the F-RAN. We name it Massive MIMO based F-RAN. For this new structure, a corresponding partially decentralized signal processing and optimization approach is proposed, in which only partial CSI is needed by the BBU pool in the cloud. Each eRRH just estimates the local CSI, with which the signals are further processed in a distributed manner. The CSI exchange between eRRHs is thus not necessary. With its limited signal processing capability resulting from the fog computing, each eRRH can perform certain tasks, which can reduce the computational burden on the BBU pool. Moreover, as we are going to show, such a design can even save hardware costs of the network. We also prove that the proposed mechanism is scalable, as the complexity is not dependent on the number of equipped antennas. Hence, increasing the number of antennas for better performance will not increase the computational complexity as well as the amount of overhead.

At the end of each Chapter, we summarize the contents and the contributions for this chapter, and give some insights and outlook for possible investigation directions in future.

In the last chapter, we conclude our work.

1.7 Related Publications and Copyright Information

As a cumulative dissertation, we would like to emphasize that several parts of this work have already been published in [CK16a; CK16b; CK16c; CK16d; CK16e; CK16f; CSK16; CK17a; CK17b; CK17c; Che18; Che+18], and these publications have been listed in Section **Publications** on page xv.

These parts, up to some modifications, are identical to the above-mentioned publications. Hence, they are ©IEEE or ©VDE. We also enrich the content with more intensive investigations that are not published yet, as well as more supportive simulation results. At the beginning of each chapter, we will clearly indicate which publications are covered within this chapter.

Chapter 2

Preliminary Information

This chapter contains

In the chapter, we are going to introduce some mathematical preliminaries in order to facilitate the future understandings and derivations.

2.1 The Rate Distortion Theory

The rate distortion theory was founded by Claude Shannon in his pioneering work on the information theory [Sha48], it provides the theoretical foundation for the lossy data compression. It determines the minimal number of bits per symbol, denoted by rate R, which should be transmitted over a channel, such that the original signal can be reconstructed at the receiver side without exceeding a given distortion metric D.

Applications in this work: As introduced previously in Chapter 1, the compression procedure plays an important role in both uplink and downlink of C-RAN/F-RAN: In the uplink, the superposed signals from all UEs at each RRH/eRRH must be compressed, before being sent to the BBU pool for further process, as the fronthaul capacity is limited. In the downlink, when the soft transfer mode is adopted, the contents intended for different UEs would be precoded, multiplexed, and modulated at the BBU pool, the resultant signals are then compressed before being sent to

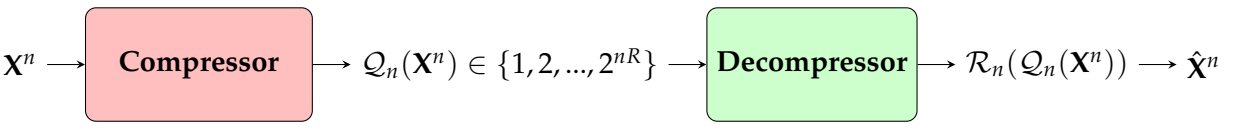


Figure 2.1: The rate distortion flow chart.

RRHs/eRRHs via fronthauls. Therefore, the compression step has to be optimized to increase the overall performance. The rate distortion theory gives the guideline in terms of how to compress, from the perspective of the information theory.

2.1.1 Definitions

Based on the definitions in [CT91], we assume that there is a source producing a sequence $X_1, X_2, ..., X_n$ i.i.d.$\sim$ $p(x), x \in \mathcal{X}$. An encoder, acting as a compressor, encodes (denoted by function $\mathcal{Q}_n()$) the source sequence $\mathbf{X}^n = \{X_1, X_2, ..., X_n\}$ into an index $\mathcal{Q}_n(\mathbf{X}^n) \in \{1, 2, 3, ..., 2^{nR}\}$. The compression index of rate R is transmitted over the channel. At the destination side, a decoder, acting as a decompressor, will decompress the the received compression index, and based on which reconstruct (denoted by function $\mathcal{R}_n()$) the original sequence. We denote the reconstructed sequence as $\hat{\mathbf{X}}^n \in \hat{\mathcal{X}}^n$. The procedure above is illustrated in Fig. 2.1.

Definition: The measure of the distortion d between the original alphabet and the reconstructed alphabet is a mapping

$$d : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}^+, \tag{2.1}$$

which is a mapping from the set of source-reconstruction alphabet pairs into the set of non-negative real numbers. The distortion $d(X, \hat{X})$ denotes the measurement between the original symbol X and the reconstructed symbol $\hat{X}$.

Definition: A distortion measurement is claimed to be bounded, if the maximal distortion value is finite, i.e.,

$$d_{\max} = \max_{X \in \mathcal{X}, \hat{X} \in \hat{\mathcal{X}}} d(X, \hat{X}) < \infty, \tag{2.2}$$

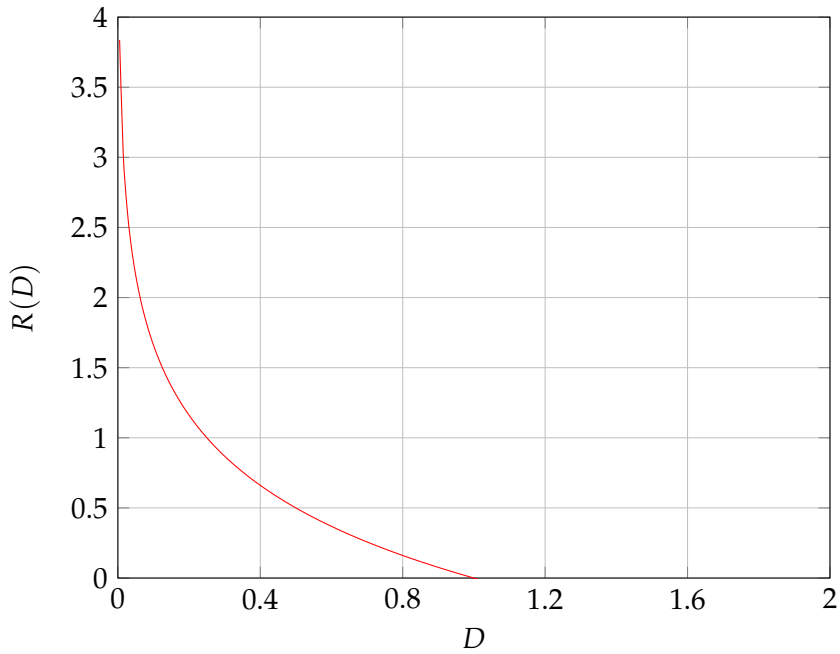
Definition: The distortion between sequence $\mathbf{X}^n$ and sequence $\hat{\mathbf{X}}^n$ is defined by

$$d(\mathbf{X}^n, \hat{\mathbf{X}}^n) = \frac{1}{n} \sum_{i=1}^{n} d(X_i, \hat{X}_i), \tag{2.3}$$

Definition: A $(2^{nR}, n)$ rate distortion code of rate R consists of the following encoding (compression) function,

$$\mathcal{Q}_n : \mathcal{X}^n \to \{1, 2, ..., 2^{nR}\}, \tag{2.4}$$

Figure 2.2: The rate distortion function of a Gaussian distributed source with mean squared error distortion.



and a decoding (reconstruction) function,

$$\mathcal{R}_n : \{1, 2, ..., 2^{nR}\} \rightarrow \hat{\mathcal{X}}^n. \tag{2.5}$$

The distortion D associated with this code, or equivalently, this compression method is defined as

$$D = \mathbb{E}\{d(\mathbf{X}^n, \mathcal{R}_n(\mathcal{Q}_n(\mathbf{X}^n)))\}. \tag{2.6}$$

A rate distortion pair (R, D) is claimed to be achievable, if there exists a code $(\mathcal{Q}_n, \mathcal{R}_n)$, or equivalently, a compression and decompression method, such that

$$\lim_{n\rightarrow\infty} \mathbb{E}\{d(\mathbf{X}^n, \mathcal{R}_n(\mathcal{Q}_n(\mathbf{X}^n)))\} \leq D. \tag{2.7}$$

The rate distortion region for a source is the closure of the set of achievable pairs (R, D). The rate distortion function $R(D)$ denotes the infimum of rates R, such that (R, D) is in the rate distortion region of the source, for a given distortion measurement D.

2.1.2 The Rate Distortion Function of a Gaussian Source

Based on the definitions above, it has been demonstrated in [CT91], that for a Gaussian source, i.e., $X \sim \mathcal{N}(0, \sigma^2)$, with squared error distortion, the rate distortion

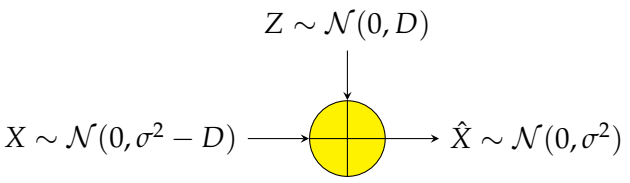


Figure 2.3: The Gaussian test channel.

function $R(D)$ can be formulated as follows and depicted in Fig. 2.2.

$$R(D) = \begin{cases} \frac{1}{2} \log \frac{\sigma^2}{D}, & 0 \leq D \leq \sigma^2, \\ 0, & D > \sigma^2. \end{cases} \tag{2.8}$$

If the scenario of communicating the source information via a capacity-limited channel is considered, it is straightforward from the figure that when higher distortion level at the destination can be tolerated, i.e., the value of D becomes larger, the code rate R can always be reduced, meaning that the source sequence can be compressed further so as to accommodate to possible worse channel qualities.

In this case, as derived in [CT91], the relationship between the original symbol X and the reconstructed symbol $\hat{X}$, or equivalently the conditional probability $p(\hat{X}|X)$, is as if the reconstructed symbol passes through a channel with additive white Gaussian noise Z of variance D, i.e., $Z \sim \mathcal{N}(0, D)$, as shown in Fig. 2.3. Due to such a relationship and the simple analytical expression of $R(D)$ in (2.8), in many existing works, the compression-decompression procedure is modeled by assuming the source signal passes through a *test channel* with additive white Gaussian noise. Such Gaussian noise acts as the distortion resulting from the compression. Then the compressor is designed based on this simple model and the resultant system is analysed from information theoretical point of view.

2.2 The Information Bottleneck Method

As described in the section above, the rate distortion theory reveals the relationship between the minimal achievable compression rate and the tolerable distortion, from the perspective of the information theory. However, in practice, the source information might have arbitrary distributions, instead of only Gaussian distribution. For a specific distribution, it is usually rather difficult to obtain the analytical expression of the rate distortion function $R(D)$. Moreover, the rate distortion theory does not directly indicate how shall the code be constructed, or in other words, how shall the compression procedure be designed, such that the rate of the compression indices is minimized.

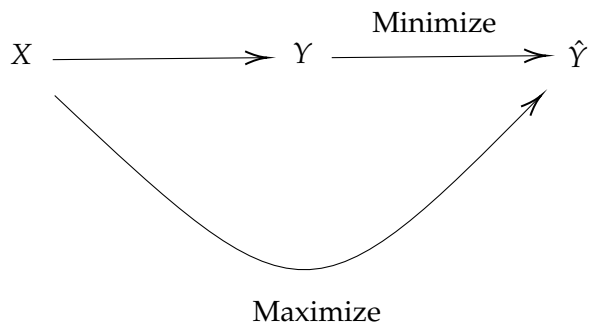


Figure 2.4: The information flow chart of the IB method.

In practice, a more meaningful concern is that, for a given rate of compression indices that can be supported by the channel, after the compression, instead of the distortion, how much ***relevant information*** between the source information and the reconstructed information at the destination can still be preserved. More specifically, in the network model considered in this work, the uplink information flow can be depicted in Fig. 2.4: After the source information denoted by X with arbitrary distribution is sent by the UE, it is distorted by the wireless channel and observed at the RRH, denoted by Y. As described in Subsection 1.2.1, the observed Y should be compressed before being delivered to the cloud via the fronthaul. Hence, Y is compressed via the compressor into a discrete compression index $\hat{Y}$, then with Wyner-Ziv coding, the forwarding rate of $\hat{Y}$ can be further reduced by generating the binning index [1]. At the BBU pool in the cloud, the source information X is going to be retrieved based on the received binning index. According to the rate distortion theory [WZ76; RHL13], the original information preserved at the BBU pool can be expressed as $I(X;\hat{Y})$. Obviously, in order to maximize the uplink transmission rate, $I(X;\hat{Y})$ shall be maximized. Furthermore, we would like to minimize $I(Y;\hat{Y})$, as it represents the transmission rate of the binning index [WZ76; RHL13]. The higher this value is, the more fronthaul capacity is required to deliver the binning index to the cloud. We also would like to obtain an analytical expression of the conditional probability $p(\hat{y}|y)$, as it directly reveals how the compressor shall be designed for the compression step.

In order to investigate the relationship between the minimized $I(\hat{Y};Y)$ and the maximized $I(X;\hat{Y})$, as well as the corresponding compression scheme, $p(\hat{y}|y)$, N. Tishby etc. proposed the so-called **Information Bottle (IB) method** in [TPB99]. The IB method is actually a special case of the rate distortion theory, such that the Kullback-Leibler divergence is adopted as the measurement of the distortion. Based on this method, the maximized $I(X;\hat{Y})$ can be computed as a function of the minimal com-

[1] For the case of a single-compressor described up to now, as no side information is available, binning has no effect on the compression rate, i.e., the transmission rate of the binning index is the same as that of the compression index. But for the case of multiple-compression, e.g., C-RAN or F-RAN, which will be introduced later, the binning can further reduce the compression rates.

pression rate, i,e., minimized $I(Y;\hat{Y})$. In detail, the function

$$I(c) = \sup_{I(Y;\hat{Y})\leq c} I(X;\hat{Y}) \tag{2.9}$$

can be computed and plotted. Hence, for a specific value of the compression rate c, whose transmission can be supported by a channel, the corresponding maximized mutual information $I(X;\hat{Y})$, denoted by $I(c)$, can be numerically obtained via the IB method. Symbol Y can be with an arbitrary distribution. Moreover, with the IB method, the way to optimally compress the source signal, i.e., the conditional probability $p(\hat{y}|y)$ that achieves compression rate c and the maximized $I(X;\hat{Y})$, can also be derived. Hence, the IB method is a powerful practical tool for the compressor design.

The relevant mutual information $I(c)$ has been proved to be a concave and increasing function for the optimized compression rate $c \in [0, H(\hat{Y})]$ [TPB99], an example is illustrated in Fig. 2.5. The IB method is a *deterministic annealing* approach such that the whole curve $I(c)$ is obtained through a third parameter $\beta, \beta > 0$, where $1/\beta = \frac{dI(c)}{dc}$ corresponds to the slope of the curve at the point $(c, I(c))$. Actually β is the *Lagrange Multiplier* used for the optimization. With the IB method, the following functional with respect to the conditional distribution is minimized:

$$\min_{p(\hat{y}|y)} I(Y;\hat{Y}) - \beta I(X;\hat{Y}). \tag{2.10}$$

We call β the trade-off factor between the compression rate c and the objective mutual information $I(c)$. By selecting an arbitrary value of $\beta > 0$ as the input of the IB method, the point on the trade-off curve with slope $1/\beta$ can be obtained. Before we briefly introduce how the IB method works, we firstly define the well-known Kullback-Leibler divergence $D_{\mathrm{KL}}(\cdot||\cdot)$ [RHL13] here: For discrete probability distributions P and Q, the Kullback-Leibler divergence between them is computed as follows

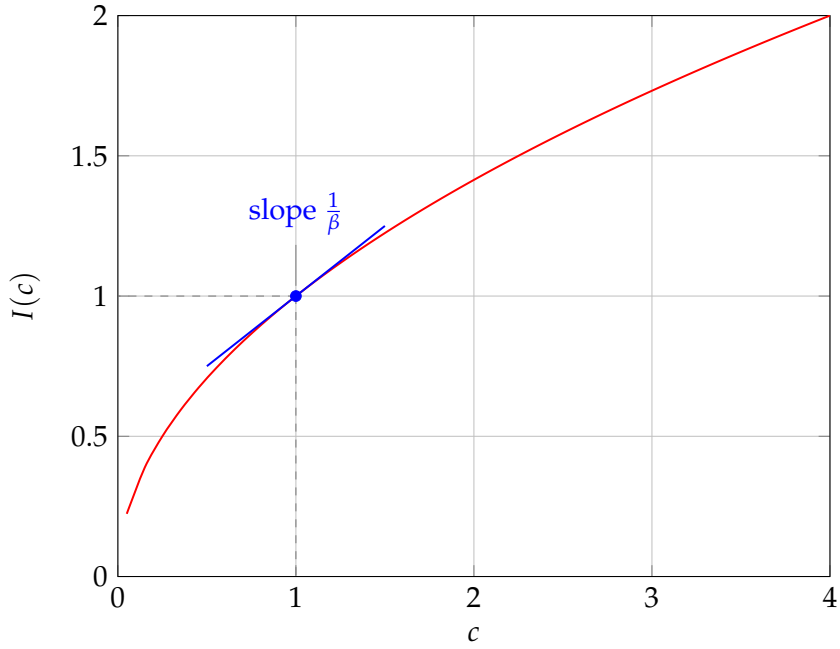
$$D_{\mathrm{KL}}(P||Q) = \sum_{x\in\mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right), \tag{2.11}$$

which is used as the distortion measurement to update the compression strategy in each iteration of the IB method.

Briefly, the IB method works in the following iterative way:

1. Select a value $\beta > 0$ as input to the IB method, and a valid initial compression strategy $p(\hat{y}|y)$;

Figure 2.5: An illustration of $I(c)$ obtained via the IB method.



2. With the compression strategy $p(\hat{y}|y)$, as well as the known channel description $p(y|x)$, we are able to compute the Kullback-Leibler divergence $D_{KL}(p(x|y)||p(x|\hat{y}))$. Then together with the input value β, a "distance" value $d(D_{KL}(p(x|y)||p(x|\hat{y})), \beta)$ shall be then computed, the analytical expression for computing d can be found in [TPB99];

3. Use the computed value d to update the compression strategy $p(\hat{y}|y)$. The updating rule is also derived in [TPB99];

4. Compute the difference between the updated $p(\hat{y}|y)$ with the one from the last iteration. If it does not converge, go to step 2. Otherwise terminate the method with the optimized $p(\hat{y}|y)$.

Therefore, with the IB method, for any specific value of $\beta > 0$, we can iteratively obtain the corresponding optimized compression strategy $p(\hat{y}|y)$. With $p(\hat{y}|y)$, the corresponding compression rate $I(Y;\hat{Y})$, denoted by c, and the *maximized* mutual information $I(X;\hat{Y})$ at the receiver side, denote by $I(c)$, can also be computed. The optimized trade-off curve Fig. 2.5 consists of different values of c and $I(c)$. As stated in [TPB99], $I(c)$ is an increasing and concave function of c. Since both values of c and $I(c)$ are monotonically increased with the value of β, by ranging the value of β from 0 to infinity as the input, the whole trade-off curve can be acquired by running the IB method accordingly. In other words, we can say that the output of the IB method for any specific value of $\beta > 0$ consists of two parts: The first part is the optimized compression strategy, i.e., $p(\hat{y}|y)$, with which the mutual information $I(Y;\hat{Y})$, i.e., the compression rate c, is minimized. Moreover, the corresponding

$I(X;\hat{Y})$ can be maximized with $p(\hat{y}|y)$, i.e., as much information is preserved at the receiver side as possible. The second part is the exact value of the compression rate $I(Y;\hat{Y})$, and the relevant information $I(X;\hat{Y})$, which are corresponding to the optimized compression strategy. Hence, if the channel capacity C is known, the compressor $p(\hat{y}|y)$ shall be designed by the IB method, such that the corresponding compression rate $c = I(Y;\hat{Y})$ is as close to C as possible, in order to fully exploit the channel resource, and preserve as much relevant information at the receiver side as possible.

Note that the IB method can generate the optimal trade-off curve by inputting different values of β , as shown in Fig. 2.5. When the compressor is designed, the location of specific points on the curve should be known, as each point on the curve corresponds to a specific optimized compression strategy, as well as the resultant compression rate and the relevant information. Hence, in order to locate a specific point (usually the point whose x-coordinate equals to the channel capacity), the Bi-Section method shall be combined with the IB method. As each value of β corresponds to a specific point on the trade-off curve, we can use the Bi-Section method to search for a specific value of β, such that at this point, the compression rate $I(Y;\hat{Y})$ can be exactly supported by the channel with capacity C, and the objective mutual information is maximized. After locating the value of β, the corresponding optimal compressor $p(\hat{y}|y)$ at this point can be acquired. Briefly, in order to find an optimal compression strategy for a channel with capacity C, the following steps shall be executed:

1. Set $\beta_L = 0, \beta_U = 100$ [2], compute $\beta = (\beta_L + \beta_U)/2$, execute the IB method with input value β.
2. Compute the compression rate corresponding to β, i.e., $I(Y;\hat{Y})$. If $I(Y;\hat{Y}) < C$, set $\beta_L = \beta$. Otherwise set $\beta_U = \beta$.
3. Update the value of β with $\beta = (\beta_L + \beta_U)/2$.
4. As long as $\beta - \beta_L > \epsilon$ is fulfilled, where ϵ is a predetermined tolerance parameter for terminating the Bi-Section method, go to step 2 to execute the IB method with the new value of β. Otherwise the searching procedure shall be terminated, the value of β is located successfully and its corresponding compression strategy is said to be optimized, with which the channel resource can be fully exploited and the relevant information is maximized.

More details of the IB method, as well as the proofs and its convergence analysis are addressed in [TPB99].

[2] Value 100 is just an example for the upper bound for the Bi-Section search here. For different scenarios in practice, different upper bounds need to be set.

Some examples of the IB method can be found in [Zei11] and [Win14]. In [Zei11], it is adopted to design the CF compressor so as to maximize the achievable rate, for a classical three-node relay network and a Multiple Access Relay Channel (MARC). In [Win14], the authors combine the Network Coding (NC) with CF, and utilize the IB method to optimize the compression process and the network encoding scheme.

Although there are various methods to realize the compression, within this work, we adopt the widely used **quantization** scheme to achieve such a compression procedure. Hence, we do not distinguish between quantization and compression in this thesis.

Applications in this work: Mainly in Chapter 3. In the uplink of C-RAN and F-RAN, the compression strategy for the superposed signals from UEs at each RRH/eRRH has to be optimized. However, as the superposed signals between RRHs/eRRHs are correlated with each other, the IB method will be extended to a so-called Alternating IB method, so as to exploit the correlation for further improving the performance. With such an extension, the compression strategies among all RRHs/eRRHs can be jointly optimized for C-RAN/F-RAN.

2.3 Optimization Techniques and Tools

The wireless system in practice is usually rather complicated. The abstracted problems resulting from the systems are non-convex in most cases. In order to investigate the design and the optimization of the network, some approximation methods and simplification schemes are widely used and have demonstrated good results. Hence, in this section, we introduce some optimization concepts, techniques and tools that will be adopted in the following chapters.

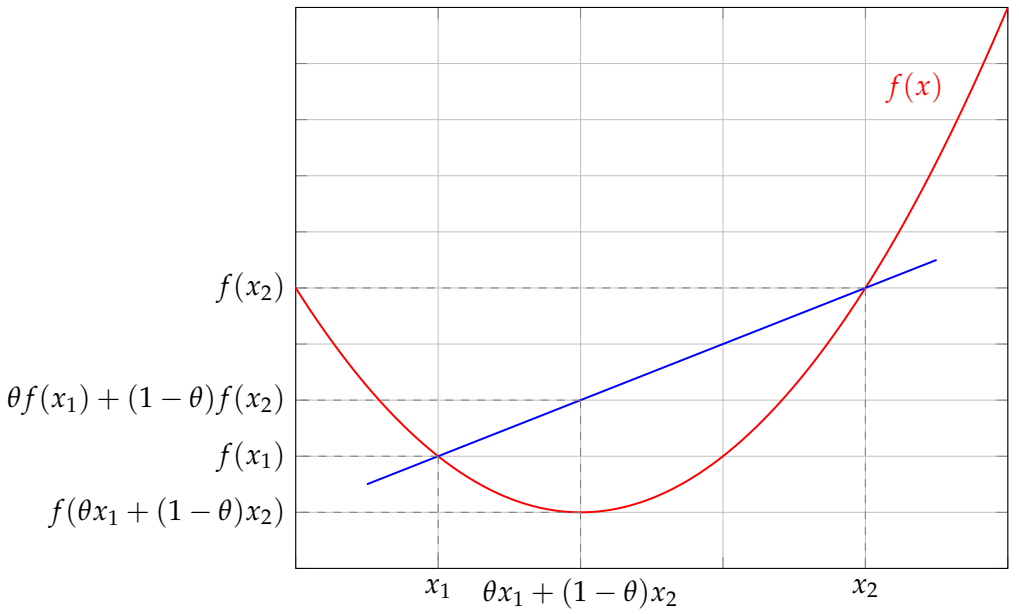
2.3.1 Convex Optimization

With the convex optimization, the solving procedures for minimizing convex functions over convex sets [BV04], is addressed. In general, a convex problem has the following form

$$\begin{aligned} &\text{minimize } f_0(\mathbf{x}), \\ &\text{subject to } f_i(\mathbf{x}) \leq b_i, i = [1:M], \end{aligned} \tag{2.12}$$

where vector $\mathbf{x} = \{x_1, x_2, ..., x_N\}$ denotes the variables to be optimized in this problem. Function $f_0 : \mathbb{R}^n \rightarrow \mathbb{R}$ denotes the objective function, which is to be minimized. Functions $f_i : \mathbb{R}^n \rightarrow \mathbb{R}$ $\forall i$ denote M constraints. The objective function

Figure 2.6: A convex function in the 2-dimensional space.



and all constraints are convex, thus they form a convex set. A vector $\mathbf{x}_{\text{opt}}$ is said to be optimal, if for any other vector $\mathbf{x}^*$, which can satisfy all constraints, inequality $f_0(\mathbf{x}^*) \geq f_0(\mathbf{x}_{\text{opt}})$ always holds.

Mathematically, a convex function indicates to a real-valued function, that is defined in an n-dimensional ($n \geq 2$) space, where the line segment between any two points of the function, does not lie under the graph. More specifically, let $f(\mathbf{x})$ be a convex function in space $\mathbb{R}^n$, $\mathbf{x}_1 \in \mathbb{R}^n$ and $\mathbf{x}_2 \in \mathbb{R}^n$ denote arbitrary two points in this space, then the following inequality must always hold:

$$f(\theta \mathbf{x}_1 + (1-\theta)\mathbf{x}_2) \leq \theta f(\mathbf{x}_1) + (1-\theta) f(\mathbf{x}_2), \tag{2.13}$$

where the value of $\theta \in [0,1] \in \mathbb{R}$ can be arbitrarily selected [BV04]. A graphic illustration of a convex function in the 2-dimensional space is depicted in Fig. 2.6.

When an optimization problem is shown to be convex, there are already sufficient algorithms, methods and tools to solve it. Many of them are explained in [BV04]. Furthermore, with MATLAB, there is a useful tool called CVX [Res20], with which the convex problems can be solved rather efficiently. By adopting CVX, MATLAB can be turned into a modeling language. For more details, please refer to [Res20]. In this work, most of the simulation results are acquired with the help of CVX. However, the systems studied in practice are usually rather complicated and not so idealized. Thus, they are mostly non-convex. In order to investigate such scenarios with many existing tools and algorithms, simplification, relaxation and approximation techniques are necessary, with which the original non-convex problem can be convexified. These techniques must be carefully designed, such that the

relaxed, or the approximated version, can still capture the essential property of the original problem. In the following parts, several widely-used convex optimization techniques will be introduced, as well as some approximation methods which will be adopted in future chapters of this thesis.

Applications in this work: The convex optimization will be utilized almost everywhere in the coming chapters. For example, in Subsection 4.2.1 of Chapter 4, the total power consumption in the downlink of F-RAN is to be minimized, in the problem, the objective is the sum of all transmission power among all eRRHs, and the constraints consist of the fronthaul capacity limitations, individual power limitations, as well as the QoS targets. As the resultant problem is non-convex, several approximation techniques are adopted to convexify the problem, then it can be solved by CVX.

2.3.2 The Bi-Section Method

The Bi-Section method is a root-finding method, i.e., locating the value of $\tilde{x}$ where $f(\tilde{x}) = 0$ holds. It bisects an interval in a repeated and iterative manner. In each step, a smaller sub-interval will be selected, where the root must be positioned.

More specifically, suppose $f(x)$ is a continuous function in interval $[a, b]$, where the signs of $f(a)$ and $f(b)$ are opposite. Hence, according to the *intermediate value theorem*, there must be at least one zero crossing within this interval, i.e., at least one specific $\tilde{x}$ exists, such that $f(\tilde{x}) = 0$ can be satisfied. The Bi-Section method is a useful tool to approach the location of this point. At the beginning, we must define a tolerance factor ϵ for terminating the procedure, this factor also indicates how precise we would like to achieve with this method. The smaller the value of ϵ is, more iterations are needed, but the obtained result would be more close to the theoretical value.

Briefly, the following steps are executed sequentially in each iteration:

1 Calculate and update the midpoint c of the current interval, i.e., $c = \frac{a+b}{2}$. Then verify whether $|c - a| > \epsilon$. If yes, proceed to the next step. If not, terminate the procedure and return $\tilde{x} = c$.

2 Calculate the value of $f(c)$.

3 If the sign of $f(a)$ and $f(c)$ are same, update the value of a by setting $a = c$, otherwise update the value of b by setting $b = c$. Hence, the searching interval for the next iteration is updated.

4 Repeat the above steps until it terminates.

Applications in this work: In this dissertation, it is used in combination with the IB method, as well as the proposed Alternating IB method, in order to locate a specific point on the trade-off curve/surface, from which the optimal compressors for the uplink of C-RAN/F-RAN can be obtained.

2.3.3 The Sub-Gradient Method

The sub-gradient method is an iterative method for solving convex problems, whose convergent behaviour is proved.

Assume that the objective function $f_0(\mathbf{x})$ is convex and we would like to obtain the optimal point $\mathbf{x}_{\text{opt}}$ which minimizes $f_0(\mathbf{x})$. With the sub-gradient method, an arbitrary valid starting $\mathbf{x}(\text{start})$ is selected firstly, i.e., $\mathbf{x}(0) = \mathbf{x}(\text{start})$, where 0 indicates that this is the initial value before the iteration procedure starts. Then in each iteration, $\mathbf{x}$ is updated as follows:

$$\mathbf{x}(\ell+1) = \mathbf{x}(\ell) - \Delta(\ell)\mathbf{g}(\ell), \tag{2.14}$$

where ℓ indicates the iteration index, and $\Delta(\ell)$ denotes the step size for this iteration. In particular, $\mathbf{g}(\ell)$ denotes the sub-gradient of $f_0(\mathbf{x})$ at point $\mathbf{x}(\ell)$. When $f_0(\mathbf{x}(\ell))$ is differentiable, $\mathbf{g}(\ell)$ is actually the gradient vector ∇f_0 at this point. Moreover, a list shall be maintained and updated in each iteration as follows:

$$f_0^{\text{opt}}(\ell) = \min\{f_0^{\text{opt}}(\ell-1), f_0(\mathbf{x}(\ell))\}. \tag{2.15}$$

Actually, it denotes the optimal value (minimized value of the objective function) we have found so far in all previous iterations. According to (2.14), in each iteration step, besides the sub-gradient $\mathbf{g}(\ell)$, the value of the step size $\Delta(\ell)$ shall also be determined and updated. There are various types of step-size determination rules whose convergences are proved, as shown in [Ber15]. Fortunately, all these rules can be determined *off-line*, i.e., before the iteration procedure starts. In this thesis, we adopt the constant step size for simplicity, i.e., $\Delta(\ell) = \Delta\ \forall \ell$.

Note that in the sub-gradient method introduced above, no constraints are assumed to exist. For the more general case where several convex constraints exist, an extension of the sub-gradient method, i.e., the *projected* sub-gradient method [BV04], has to be applied. The convex constraints in (2.12) can be denoted by

$$\mathbf{x} \in \mathcal{C},$$

where $\mathcal{C}$ represents the *convex set* described by all convex constraints. Then the updating rule for each iteration becomes

$$\mathbf{x}(\ell+1) = \Pi_{\mathcal{C}}\left(\mathbf{x}(\ell) - \Delta(\ell)\mathbf{g}(\ell)\right), \tag{2.16}$$

where $= \Pi_{\mathcal{C}}()$ indicates the projection on $\mathcal{C}$. More details of these methods are documented in [BV04].

Applications in this work: The sub-gradient method is utilized to optimize the power allocation in the downlink of F-RAN in Subsection 4.2.4, where the network multi-cast throughput is to be maximized.

2.3.4 The Semi-Definite Relaxation (SDR)

The Semi-Definite Relaxation (SDR) is a relaxation technique that can convert a non-convex problem into a Semi-Definite Programming (SDP) problem, its effectiveness as well as the correctness has been deeply studied by many works. A SDP problem can be efficiently solved by many existing tools, e.g., CVX. Such a technique is widely used in the field of signal processing and wireless communication, e.g., the problem of MIMO detection and transmit beamforming.

Suppose we have a problem with the following formulation

$$\begin{aligned} &\min_{\mathbf{x}\in\mathbb{R}^N} \mathbf{x}^T\mathbf{C}\mathbf{x}, \\ &\text{subject to } \mathbf{x}^T\mathbf{A}_i\mathbf{x} \leq b_i, i = [1:M], \end{aligned} \tag{2.17}$$

where $\mathbf{C}$ and $\mathbf{A}_i$ $\forall i$ are all positive semi-definite matrices, i.e., $\mathbf{C}, \mathbf{A}_i \succeq \mathbf{0}$ $\forall i$, and $\mathbf{x} \in \mathbb{R}^{N\times 1}$ is the vector of variables that needs to be optimized. This problem is non-convex and NP-hard.

In order to solve this problem, the SDR is an effective tool for convexification and simplification. Note that we can express $\mathbf{x}^T\mathbf{C}\mathbf{x} = \mathrm{Tr}(\mathbf{x}^T\mathbf{C}\mathbf{x}) = \mathrm{Tr}(\mathbf{C}\mathbf{x}\mathbf{x}^T)$ where $\mathrm{Tr}(\cdot)$ denotes trace of a matrix. Then by noting that $\mathbf{X} = \mathbf{x}\mathbf{x}^T$, we have $\mathbf{x}^T\mathbf{C}\mathbf{x} = \mathrm{Tr}(\mathbf{C}\mathbf{X})$.

Hence, the original problem (2.17) can be equivalently reformulated as follows

$$\begin{aligned} &\min_{\mathbf{x}\in\mathbb{R}^N} \mathrm{Tr}(\mathbf{C}\mathbf{X}), \\ &\text{subject to } \mathrm{Tr}(\mathbf{A}_i\mathbf{X}) \leq b_i, i = [1:M], \\ &\mathbf{X} \succeq \mathbf{0}, \\ &\mathrm{rank}(\mathbf{X}) = 1. \end{aligned} \tag{2.18}$$

The objective function, as well as the first two constraints are convex. However, the last constraint, i.e., $\text{rank}(\mathbf{X}) = 1$, makes the problem above non-convex and NP-hard. The basic idea of SDR is to relax the problem by dropping the last constraint, i.e., the rank limitation. Then the remaining objective and constraints would form a convex SDP problem, which can be solved by many existing efficient methods. After the solution of the relaxed SDP problem is acquired, it shall be converted to an approximated solution of the original problem, by involving the rank limitation again, which has been omitted in the relaxation procedure. This can be done with, e.g., the EigenValue Decomposition (EVD) method or the randomization and scaling method. The final solution is naturally sub-optimal. In [KSL08] and [Luo+10], this technique and its application to signal processing and wireless communication are intensively introduced and investigated.

Applications in this work: The SDR technique plays an important role in this dissertation. As already introduced in Chapter 1, from the viewpoint of the BBU pool, the C-RAN/F-RAN can be regarded as a networked MIMO system. Hence, when the aggregated beamformers/precoders are to be designed, the SDR technique is adopted to convexify and relax the original problem. As we are going to see in Chapter 4, the SDR technique appears in both high EE and SE oriented design, as well as the robust design of the network when only inaccurate CSI is available.

2.3.5 ℓ_0-norm Approximation

In many scenarios, we have to deal with optimization problems with discrete objective functions. In the downlink design of F-RAN for example, when how to cluster different RRHs/eRRHs to serve multi-cast groups optimally is investigated, clusters consisting of different sets of RRHs/eRRHs are obviously discrete functions, and thus, non-convex. Such problems are called Mixed Integer Non-Linear Programming (MINLP) problems [MFR20], which are NP-hard.

First of all, we introduce the mathematical definition of a norm for future investigation. The ℓ_p-norm ($p \geq 1$) of a vector $\mathbf{x} = (x_1, x_2, ..., x_N)$, i.e., $||\mathbf{x}||_p$ is defined as follows:

$$||\mathbf{x}||_p := \left(\sum_{i=1}^{N} |x_i|^p \right)^{1/p}. \tag{2.19}$$

Specifically, the ℓ_0-norm of $\mathbf{x}$ is defined as an indicator to the number of non-zero elements in the vector, i.e.,

$$|\mathbf{x}|_0 = \#(x_i) \text{ with } x_i \neq 0. \tag{2.20}$$

For problems we address in later chapters, we can equivalently rewrite the discrete MINLP problem into the following form:

$$\begin{aligned} &\min_{\mathbf{x}\in\mathbb{R}^N} \sum_{i=1}^{N} a_i|x_i|_0 + f_0(\mathbf{x}), \\ &\text{subject to } \sum_{i=1}^{N} b_{i,j}|x_i|_0 + f_j(\mathbf{x}) \leq c_j, j = [1:M], \end{aligned} \tag{2.21}$$

where $f_j(\mathbf{x})\ \forall j \in [0:M]$ are convex functions, and $a_i, b_{i,j}, c_j\ \forall i, j$ are real constant values. Obviously, the ℓ_0-norms make the problem discrete, non-convex and NP-hard. The technique to tackle such problems is an iterative ℓ_0-norm approximation method, which is widely used in the field of Compressed Sensing [CWB08]. In this method, the discrete ℓ_0-norm is approximated by a linear function of it. And in each approximation iteration, the coefficient of this linear function is recalculated and updated. More specifically, the ℓ_0-norm of x_i is iteratively approximated as

$$|x_i^{(t+1)}|_0 \approx w_i^{(t+1)} x_i^{(t+1)}, \text{with } w_i^{(t+1)} = \frac{1}{x_i^{(t)} + \tau}, \tag{2.22}$$

where t denotes the iteration index, w_i is called the re-weighted coefficient of x_i, and τ is the threshold parameter that shall be determined in advance by us, according to the actual situation and the target we would like to achieve. With such an approximation, the discrete non-convex term now becomes linear and convex. In order to make it easier to follow, we firstly drop the superscript (t) and $(t+1)$ to explain such an approximation: Now we have $|x_i|_0 \approx w_i \cdot x_i = \frac{x_i}{x_i+\tau}$. When $x_i \gg \tau$ holds, the approximation of ℓ_0-norm is rather close to 1. Contrarily, the approximation would rapidly approach 0, when $x_i \ll \tau$. Therefore, τ can be regarded as a threshold parameter, which determines whether the value of x_i is turned on (1), or switched off (0). By carefully selecting the value of τ, this continuous and linear approximation can capture the behavior of discrete non-convex ℓ_0-norm.

The superscripts in (2.22) reflect the iterative re-weighted procedure. In the t-th iteration, the approximated minimization problem, which is convex according to (2.21) (since we have convexified all non-convex terms with this approximation method) can be solved, then $x^{(t)}$ will be obtained, and $w^{(t+1)}$ used for the next iteration can be computed and updated accordingly. When the value of the obtained $x_i^{(t)}$ decreases in iteration t compared to the previous iteration, it must have larger re-weighted coefficient $w_i^{(t+1)}$ in the next iteration. Hence, its value will be forced to further decrease, and be encouraged to drop below the threshold value τ. By continuing such a re-weighting procedure iteratively, some elements of $\mathbf{x}$ will be finally forced to be rather close to 0 (they can be regarded as 0 as long as the value of them fall below the value of the predetermined threshold parameter τ), and the remaining elements can still satisfy all constraints. Therefore, the NP-hard MINLP

problem can be avoided. More detailed introduction is documented in [CWB08].

Applications in this work: In the downlink of F-RAN, there are two eRRH selection issues: The first one is the cluster formulation: As the fronthaul capacity is limited, it might be not possible that all eRRHs serve for all scheduled UEs. Thus for each requested content, a subset of eRRHs shall be selected to form a cluster for its transmission. Hence, the BBU pool has to decide which eRRH shall be in which subset, via the optimization procedure. The second one is when eRRH deactivation is considered to save power, the BBU pool shall determine which eRRHs can be switched off, such that the remaining ones can still fulfill the requirements of the network. For both cases, the ℓ_0-norm is utilized to denote such a selection. Hence, the corresponding optimization requires the iterative approximation method introduced above.

2.3.6 S-Lemma

The S-Lemma [DM06] is an effective tool for tackling the problem of the robust optimization, which is widely used in many field, e.g., the control theory. In this work, we adopt this lemma to design the robust networks, in which the QoS of each UE can still be guaranteed, even with inaccurate CSI. The S-Lemma is summarized as follows:

S-Lemma: Let two functions $f_0(\mathbf{x}), f_1(\mathbf{x})$ defined as $f_0(\mathbf{x}) = \mathbf{x}^H\mathbf{A}_0\mathbf{x} + 2\text{Re}\{\mathbf{x}^H\mathbf{b}_0\} + c_0$ and $f_1(\mathbf{x}) = \mathbf{x}^H\mathbf{A}_1\mathbf{x} + 2\text{Re}\{\mathbf{x}^H\mathbf{b}_1\} + c_1$, where $\mathbf{b}_0, \mathbf{b}_1 \in \mathbb{C}^{d\times 1}$ denote vectors, matrices $\mathbf{A}_0, \mathbf{A}_1 \in \mathbb{C}^{d\times d}$ are all Hermitian matrices; and c_0, c_1 are scalars. Suppose that a specific vector $\hat{\mathbf{x}} \in \mathbb{C}^{d\times 1}$ exists, with which $f_1(\hat{\mathbf{x}}) < 0$ is satisfied. Then $f_0(\mathbf{x}) \geq 0$ and $f_1(\mathbf{x}) \leq 0$ can be satisfied simultaneously, for **arbitrary** $\mathbf{x} \in \mathbb{C}^{d\times 1}$, as long as a scalar $\lambda \geq 0$ exists, which makes the following matrix positive semi-definite, i.e.,

$$\begin{bmatrix} \mathbf{A}_0 & \mathbf{b}_0 \\ \mathbf{b}_0^H & c_0 \end{bmatrix} + \lambda \begin{bmatrix} \mathbf{A}_1 & \mathbf{b}_1 \\ \mathbf{b}_1^H & c_1 \end{bmatrix} \succeq \mathbf{0}. \tag{2.23}$$

Chapter 3

Centralized Joint Design for the Uplink

This chapter contains

[1] When we talk about something like "*5G is must faster than 4G...*", we usually mean the network throughput, i.e., the total achievable transmission rate of the network, is much faster. More specifically, with 5G, the users can experience much faster data transmission rates for both uplink and downlink. Concerning the uplink, it denotes the slots, in which the UE has the opportunity to upload their own data to the core network. For example, if we would like to share a photo with a friend via some mobile Apps in a 5G environment, after we click *send*, this photo is going to be sent from your mobile phone to the core network, via the uplink transmission. Contrarily, in the downlink slots, the UE has the opportunity to receive data from the network: The photo you have just sent, will be downloaded by your friend via the downlink transmission. The uplink and downlink are orthogonal, i.e., they share the network resources, consisting of time and frequency, in an orthogonal way. In 5G TDD (Time Division Duplex) mode, the uplink and the downlink slots arrive at different time slots. While in 5G FDD (Frequency Division Duplex) mode, they arrive at different frequency bands.

[1] Parts of this chapter have been published in [CK16b; CK16c; CK16d; CK16e].

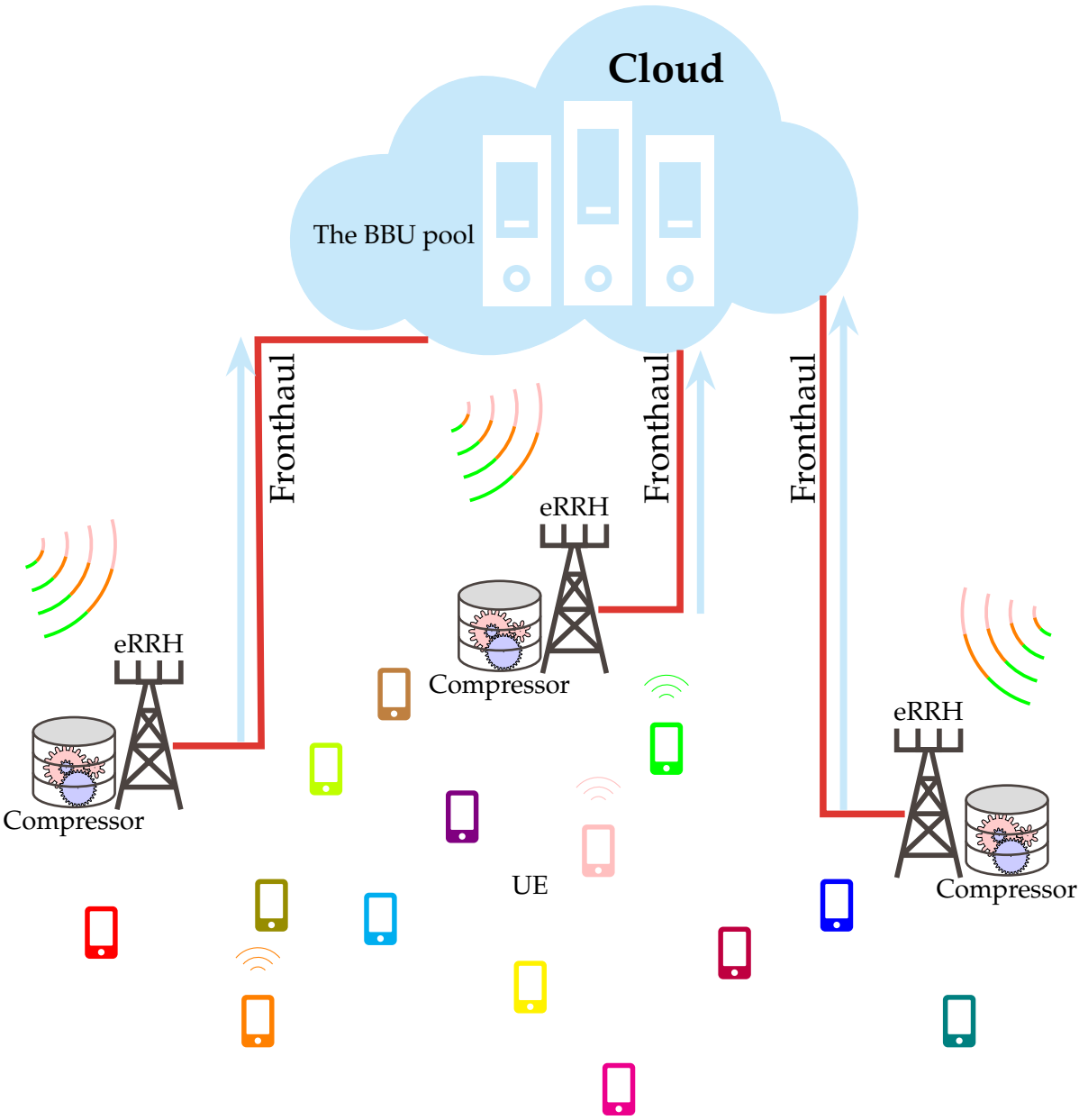


Figure 3.1: The uplink transmission of F-RAN. UEs emitting signals are scheduled in this uplink slot, and the emitted signals are superposed at each eRRH.

When it comes to the optimal design of the C-RAN/F-RAN, the uplink and downlink are two distinct stories. In this chapter, we will focus on investigating the uplink transmission of C-RAN and F-RAN, which is sketched in Fig. 3.1. As seen from the figure, in each uplink slot, some UEs will get scheduled [2], meaning that only these UEs can upload their data for this time being, i.e., within this time slot, while the others have to be silent.

After the scheduled UEs send their signals, all signals are independently superposed at each RRH/eRRH. As stated in Chapter 1, unlike the numerous tasks undertaken by the BS in 4G network, the C-RAN/F-RAN architecture adopted in 5G implements rather limited signal processing functionalities at RRHs/eRRHs, most of them are pushed to the BBU pool in the cloud. Therefore, the fronthaul, which links the RRH/eRRH and the BBU pool, has to transmit such almost *raw* signals. We have already illustrated in Fig. 1.5 how much capacity is required at fronthauls. Therefore, a large amount of hardware costs might be saved, if these signals can be somehow efficiently compressed before being sent to the BBU pool. Moreover, at the BBU pool, the compressed signals are expected to retain as much useful information as possible. Such requirements have motivated us to consider the Information Bottleneck method, introduced in Section 2.2, with which the compression strategies can be optimized. We are going to investigate the application and extension of the IB method in C-RAN/F-RAN in the coming parts of this chapter.

Obviously, in the uplink model considered in this chapter, RRH/eRRH is the only place where some compression strategies are implemented. Thanks to the concept of the fog computing, RRHs can evolve to eRRHs by being equipped with some quantizers, for the execution of the compression step. The quantizers are designed, such that the transmission of the compressed signals resultant from the quantization step, can be supported by the fronthauls. Then after the BBU pool receives the compressed signals, they shall be reconstructed in the cloud. Therefore, the design of the quantization step is critical, when the uplink transmission is to be optimized. The design of the quantizer at each eRRH is one of the main topics and contributions of this chapter.

Note that in the discussion above, we assume that the fronthaul resource for each eRRH is fixed. Thus, only the quantizers are to be optimized in order to meet the fronthaul capacity constraints. However, when several eRRHs share the capacity of fronthaul, the issue of the resource allocation on the fronthaul shall also be discussed. It can be regarded as an extension to the problem of the quantizer design introduced above, due to the interaction between the compression step, and the

[2] The notifications of which UEs are get scheduled, are usually sent several slots ago in the UL-DCI (Downlink Control Information) via PDCCH (Physical Downlink Control Channel). UL-DCI indicates who will be scheduled, as well as in which uplink slots they are scheduled. More details can be found in [3GP18].

available fronthaul capacity. In such a scenario, how much capacity shall be allocated to each eRRH will be addressed.

There are already considerable amount of works addressing the optimization of the quantizers for the uplink of C-RAN. First of all, mainly two compression strategies are investigated, i.e., Compress-and-Forward (CF) [CG79] and Noisy Network Coding (NNC) [Lim+11]. When CF is performed, the Wyner-Ziv coding [WZ76] is exploited at RRHs since the signals received by neighboring RRHs are statistically correlated. The BBU pool implements successive decompression and decoding. When NNC is utilized, the RRHs perform quantization without the Wyner-Ziv coding, and the BBU pool does simultaneous joint decompression and decoding among all received blocks. Generally, the throughput of NNC is higher than that of CF [Lim+11], while its complexity is much higher and the delay is much longer, as it requests consecutive data blocks to be received before the decoding procedure being executed, so as to approach the optimums. In many existing works, the optimization of the quantizers is studied from an information theoretical point of view by exploiting the rate distortion theory, i.e., only the Gaussian codebook adopted by users is considered, and the quantization procedure is modeled by a Gaussian test channel, see Fig. 2.3. Under such an assumption, the optimization of the quantization noise levels for these two strategies is investigated. In [ZY14], an Alternating Convex Optimization (ACO) approach is proposed for CF, and in [Par+13b], an iterative algorithm based on the Majorization Minimization (MM) approach is considered for NNC.

However, as stated in Chapter 2, the rate distortion theory, as well as the Gaussian test channel cannot instruct the quantizers' design for the F-RAN model considered this work, when arbitrary codebooks are adopted and the target is to maximize the preserved information. In the uplink of F-RAN, the quantization information flow can be described as follows: The scheduled UE can use arbitrary codebook $\mathcal{X}$ with a finite alphabet, and the received signal at each eRRH is discretized and sampled firstly into finite alphabet $\mathcal{Y}$, then based on the compression scheme described as $P_{\hat{Y}|Y}$, it will be further compressed into several quantization levels, denoted by $\hat{\mathcal{Y}}$. Usually its cardinality, i.e., $|\hat{\mathcal{Y}}|$, is much smaller than $|\mathcal{Y}|$ due to the compression. Then $\hat{Y}$ is encoded and transmitted by the fronthaul with its limited capacity. After the BBU pool decodes $\hat{Y}$, it tries to extract the useful information of each UE from it. In such a scenario, the Information Bottleneck (IB) method [TPB99] is a useful tool to optimize the quantizer $P_{\hat{Y}|Y}$, such that the trade-off between the preserved information $I(X;\hat{Y})$, and the compression rate $I(Y;\hat{Y})$ can be found. Hence, for the uplink of C-RAN and F-RAN, we consider to use the IB method to design optimal quantizers at RRHs.

However, the IB method is considered only for the case of single quantizer in most works. When multiple quantizers exist, as in C-RAN and F-RAN, the optimization

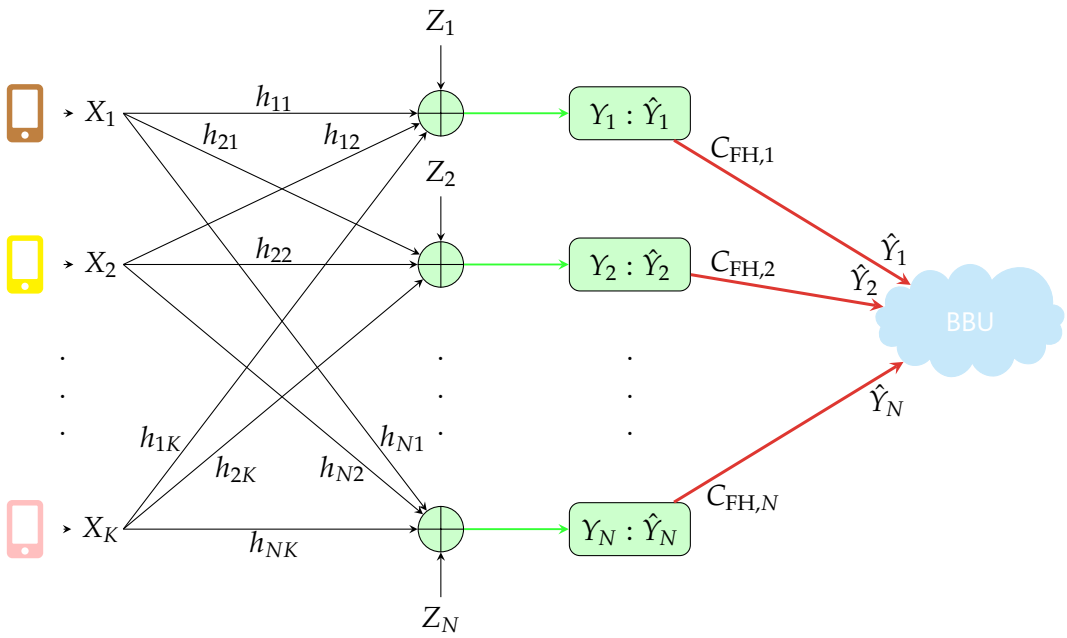


Figure 3.2: The abstract model for the uplink of F-RAN.

of the quantizers depends not only on its own received signal, but also on that at other eRRHs, as well as their compression strategies. This is mainly due to the fact that the received signals are correlated and thus the Wyner-Ziv coding is performed. Thus a joint optimization among all quantizers is required, which is difficult to be implemented in the conventional RAN. Thanks to the BBU pool with high computational capability, such a joint optimization is possible in C-RAN and F-RAN. Then the problem becomes, how to extend the conventional IB method to the case of multi-quantizer, where the quantization steps performed by them are correlated and influence with each other. In this work, we propose a so-called Alternating Information Bottleneck (AIB) method and an alternating Bi-Section method, from which all quantizers at eRRHs can be jointly optimized.

3.1 System Model

3.1.1 Overview

We consider the abstract uplink model depicted in Fig. 3.2, where UEs intent to send their data to the cloud server via the uplink transmission. eRRHs at the other side of the radio access channel observe different and independent linear combinations of the original signals plus additive white Gaussian noise. In order to accommodate to the limited fronthaul capacities, the quantizer at each eRRH compresses the superposed signal into a compression index. As the signals received by neighboring

eRRHs are correlated, Wyner-Ziv coding is adopted to further reduce the compression rates. The resultant binning indices are encoded and sent via the fronthauls. The BBU pool decodes all binning indices and performs a joint decompression and decoding, so as to extract the original message of each UE. In this work, we focus on designing compression strategies among all eRRHs such that the BBU pool is able to extract as much original information as possible. In other words, the end to end achievable sum rate of the uplink is maximized.

3.1.2 Mobile Users and Remote Radio Heads

The network is assumed to have K single-antenna UEs sending independent messages with arbitrary codebooks and modulation schemes. Totally N eRRHs acting as signal collectors are deployed within the whole network. For illustrative simplicity, we only discuss the case of single-antenna eRRHs, but the proposed algorithms and results can also be extended to the MIMO case, as we are going to show later on.

3.1.3 Radio Access Channel

Let X_k denote the transmitted symbol from the k-th UE, with arbitrary modulation scheme and power denoted by $P_k = \mathbb{E}\{|X_k|^2\}$, and h_{nk} denote the complex channel coefficient from the k-th UE to the n-th eRRH. Hence, at the n-th eRRH, the received analog signal $Y_{n,\text{analog}}$ can be expressed as

$$Y_{n,\text{analog}} = \sum_{k=1}^{K} h_{nk} X_k + Z_n,\ n \in \{1, 2, ..N\},$$

where $Z_n \sim \mathbb{C}\mathcal{N}(0, \sigma_n^2)$ is the additive white Gaussian noise with variance σ_n^2. Therefore, the radio access channel between the UEs and eRRHs is actually an $N \times K$ interference channel.

3.1.4 Compression at eRRHs

The received analog signal $Y_{n,\text{analog}}$ is first sampled and discretized into Y_n with finite alphabets $\mathcal{Y}_n$. Actually, such a discretization, or in another word, analogue-to-digital conversion, can be regarded as a pre-quantization step. Such a conversion is essential in the current digital communication system. When we talk about the compression step, which is performed by the quantizers at eRRHs, we indicate the further quantization of the discretized signal Y_n into $\hat{Y}_n$, so as to meet the fronthaul

capacity constraint. More specifically, after the received analogue signal $Y_{n,\text{analog}}$ is discretized into a digital signal Y_n, the eRRH performs compress-and-forward (CF) on it: Its quantizer compresses the signal Y_n into $\hat{Y}_n$ based on the compression scheme $P_{\hat{Y}_n|Y_n}$, which is going to be optimized in this chapter. The cardinality of the alphabet of $\hat{Y}_n$, i.e., $|\hat{\mathcal{Y}}_n|$, is assumed to be much smaller than the cardinality of the alphabet of Y_n, i.e., $|\hat{\mathcal{Y}}_n| \ll |\mathcal{Y}_n|$. Since the received signals of the neighboring eRRHs are statistically correlated, the Wyner-Ziv coding is to be utilized, with which binning indices are generated.

3.1.5 Fronthaul Transmission

The binning indices are then encoded and transmitted by fronthauls from each eRRH to the BBU pool in the cloud. The capacity of the n-th fronthaul, i.e., the fronthaul connecting the n-th eRRH and the cloud, is denoted by $C_{\text{FH},n}$ and known to the BBU pool. Moreover, as we focus on the optimization of the compression strategies, an error-free transmission of the binning indices via fronthauls is assumed, i.e., the encoded binning indices can be perfectly decoded by the BBU pool in the next step.

3.1.6 Centralized Processing in the Cloud

The global CSI is supposed to be available at the BBU pool in the cloud. The BBU pool adopts a successive two-stage decoding: It first decodes all binning indices from all eRRHs, and then decodes UEs' messages $\boldsymbol{X} = [X_1, X_2, ..., X_K]^T$ based on the decoded binning indices. Compared with the NNC, where a simultaneous joint decoding of compressed signals and the desired messages over all received blocks is required, the successive decoding nature of CF overcomes some difficulties in the practical implementation of the NNC, such as the long latency and the high computational complexity. Moreover, we assume that the modulation scheme of each UE are available at the BBU pool [3], i.e., the probability distribution of $\boldsymbol{X}$ is known, and the design of the optimized quantizers can be feed-backed to the corresponding eRRHs.

Remark: Since the Wyner-Ziv coding is utilized, the decompression order π generally affects the achievable performance and shall be optimized upon. In this work we will not address this problem. According to [Par+14], a generally sensible, and close to optimal choice is to firstly decompress the signals coming from the eRRHs

[3] This is the usual case in the uplink, as the Modulation and Coding Scheme (MCS) of each scheduled UE is actually determined at various types of 5G BS, and is also notified by the UL-DCI via PDCCH several slots before.

with larger fronthaul capacities and then those with smaller ones. The rationale is that the compressed information from the eRRH with large fronthaul capacity provides more relevant side information for the others. We adopt such an ordering in this paper. Without loss of generality, we assume $C_{\text{FH},1} \geq C_{\text{FH},2} \geq ... \geq C_{\text{FH},N}$ and the decompression ordering π is $\pi(n) = n,\ n \in \{1, 2, ..., N\}$.

3.1.7 Problem Statement

We aim to maximize the achievable sum rate [Par+14] in the uplink of F-RAN as follows:

$$\begin{aligned} &\max_{P_{\hat{Y}|Y}} I(\boldsymbol{X}; \hat{\boldsymbol{Y}}), \\ &\text{subject to } I(Y_n; \hat{Y}_n | \hat{Y}_1, ..., \hat{Y}_{n-1}) \leq C_{\text{FH},n},\ \forall n \in \{1, 2, ..., N\}, \end{aligned} \tag{3.1}$$

where $\boldsymbol{X} = \{X_1, X_2, ..., X_K\}$, $\hat{\boldsymbol{Y}} = \{\hat{Y}_1, \hat{Y}_2, ...\hat{Y}_N\}$ and $P_{\hat{\boldsymbol{Y}}|\boldsymbol{Y}} = \prod_{i=1}^{N} P_{\hat{Y}_i|Y_i}$. The mutual information between the original message $\boldsymbol{X}$ from all UEs and the obtained compressed information $\hat{\boldsymbol{Y}}$ determines how much information can be retrieved finally by the BBU pool in the cloud. The rate of the binning indices for the signal received by the n-th eRRH is $I(Y_n; \hat{Y}_n | \hat{Y}_1, ..., \hat{Y}_{n-1})$, as the Wyner-Ziv coding is utilized. Naturally, $\sum_{|\hat{y}_n|} P_{\hat{Y}_n|Y_n} = 1,\ \forall y_i$ and $P_{\hat{Y}_n|Y_n} \geq 0,\ \forall \hat{y}_n, y_n$ shall be satisfied. We see that when the modulations schemes, capacities of the fronthaul links and the channel configuration are fixed, the sum rate depends solely on how eRRHs compress their received signals.

Note that the fronthaul capacities are finite, therefore, on the one hand, the quantization cannot be too fine in the compression step, such that the compression rate might exceed the fronthaul capacity, which will lead to decoding failure of the binning indices by the BBU pool. On the other hand, if the quantization is too coarse, the capability of the fronthaul link might not be fully utilized, the overall performance is thus limited by the coarse quantization. Hence, an optimal trade-off between the compression rates and the achievable sum rate must be found. As stated in the chapter before, the Information Bottleneck (IB) method [TPB99] is an effective tool to find such a trade-off as well as the corresponding optimized compression strategy, in case of the single quantizer. In this work, we extend the conventional IB method to an alternating IB method, which is able to deal with the case of the correlated multiple quantizers in F-RAN.

3.2 Alternating Information Bottleneck Optimization

In this section, we firstly derive the AIB method and show its application to the joint optimization of the quantizers at eRRHs. Then we analyse the convergence behaviour of the AIB method. A so-called alternating Bi-Section method will also be proposed, with which a specific point on the trade-off surface can be obtained, where the optimal quantizers can be derived which can fully utilize the fronthaul resources and maximize the sum rate. Furthermore, we address the problem of resource allocation on the fronthaul with the help of the proposed algorithms. Finally the numerical results obtained via the simulation will be given.

3.2.1 The Alternating Information Bottleneck Method

For ease of the illustration of the AIB method, and the analysis of its convergence behavior, we start with the F-RAN consisting of two UEs and two eRRHs, each is equipped with a single antenna. At the end of this section, we will show that our proposed optimization algorithms can be conceptually easily extended [4] to the case of more UEs with more antennas. According to (3.1), the problem becomes

$$\begin{aligned} \max_{P_{\hat{Y}_1|Y_1} P_{\hat{Y}_2|Y_2}} \quad & I(X_1, X_2; \hat{Y}_1, \hat{Y}_2), \\ \text{subject to } & I(Y_1; \hat{Y}_1) \leq C_{\text{FH},1}, \\ & I(Y_2; \hat{Y}_2|\hat{Y}_1) \leq C_{\text{FH},2}. \end{aligned} \tag{3.2}$$

At first we set up the trade-offs between the compression rate pair of two eRRHs and the corresponding maximized sum rate. It shall be emphasized that such trade-offs form a curve or a surface, which consist of various optimal points. How to locate a specific point on it, in order to accommodate to specific fronthaul capacities will be introduced in the next section. More specifically, the constraints in (3.2) should be satisfied with equality, aiming to fully exploit the available fronthaul capacity, so as to maximize the sum rate.

In detail, we propose the AIB method, with which the trade-off between the compression rate pair $\big(I(Y_1; \hat{Y}_1), I(Y_2; \hat{Y}_2|\hat{Y}_1)\big)$, and the corresponding maximized achieved sum rate $I(X_1, X_2; \hat{Y}_1, \hat{Y}_2)$, can be derived, through the trade-off factor pair (β_1, β_2). It is also worth to mention that when the AIB method is interpreted, we select the decompression order $\hat{Y}_1 \rightarrow \hat{Y}_2$, i.e., the signal $\hat{Y}_1$ from eRRH 1 is decompressed firstly, the signal $\hat{Y}_2$ is then decompressed with $\hat{Y}_1$ as the side information. It is also possible to do it the other way around, but the derivation procedures are

[4] We say *conceptually easily extended* as the mathematical extension of the proposed algorithm is straightforward. However, the computational complexity will increase exponentially.

still the same. The optimal decompression order is out the scope of this thesis, but in Subsection 3.4.1 and 3.5.1, more comments on the decompression order will be given, and the results between different orders will be compared.

As we have mentioned before, in C-RAN/F-RAN, as the quantizers at eRRHs need to be optimized jointly, the conventional IB method cannot be adopted directly. However, in the two-eRRH scenario considered here, we notice that if one quantizer is fixed, the remaining part is simply in a form, such that the conventional IB method can be readily utilized. Basically, the main idea of the AIB method is that, different quantizers are fixed alternatively: In each alternating step, only a specific quantizer is optimized, all other compression strategies at other quantizers are fixed, except for the one that is to be optimized in the current step. Hence, the conventional IB method can be applied now, as just one quantizer is to be optimized. After the optimization, the updated compression strategy of this quantizer is fixed, and the next one is to be optimized similarly. Such an alternating step will definitely converge, as we are going to show later in this subsection.

Now we go back to the specific scenario stated above to illustrate the AIB method in detail: In the algorithms proposed below, **Function IB2** is the algorithm to optimize the compression strategy for the quantizer at eRRH 2, by assuming that the compression strategy of eRRH 1 is known and fixed. Similarly, **Function IB1** is used to optimize the compression strategy for eRRH 1, by assuming a fixed compression strategy of eRRH 2. These two functions are going to be called in different alternating steps in **Function AIB**, which is the proposed Alternating Information Bottleneck method. The AIB method is derived as follows:

1. If the first quantizer $P_{\hat{Y}_1|Y_1}$ at eRRH 1 is fixed, then the job is reduced to find the optimal trade-off between the compression rate $c_2 = I(Y_2;\hat{Y}_2|\hat{Y}_1)$ and $\max_{P_{\hat{Y}_2|Y_2}} I(X_1,X_2;\hat{Y}_1,\hat{Y}_2)$ for the quantization operation executed at eRRH 2. Due to the chain rule

$$I(X_1,X_2;\hat{Y}_1,\hat{Y}_2) = I(X_1,X_2;\hat{Y}_1) + I(X_1,X_2;\hat{Y}_2|\hat{Y}_1),$$

then it is sufficient to compute the trade-off between $I(Y_2;\hat{Y}_2|\hat{Y}_1)$ and $\max_{P_{\hat{Y}_2|Y_2}} I(X_1,X_2;\hat{Y}_2|\hat{Y}_1)$, as the value of $I(X_1,X_2;\hat{Y}_1)$ is fixed. Now it is further reduced to the problem solved in [Zei11] with the conventional IB method, as shown in **Function IB2**. In this function, the fixed quantizer at eRRH 1, i.e., $P^{\text{fixed}}_{\hat{Y}_1|Y_1}$ is the input and a local invariant when optimizing the quantizer at eRRH 2, i.e., $P_{\hat{Y}_2|Y_2}$. As already introduced in Subsection 2.2, the IB method requires an initial compression strategy, then it is iteratively updated and optimized. In **Function IB2**, $P^{\text{init}}_{\hat{Y}_2|Y_2}$ denotes such an initial mapping. Moreover, it is assumed that the CSI knowledge and

Function IB2($P^{\text{fixed}}_{\hat{Y}_1|Y_1}$, $P^{\text{init}}_{\hat{Y}_2|Y_2}$, $P_{Y_1,Y_2|X_1,X_2}$, $|\hat{\mathcal{Y}}_2|$, β_2, ϵ_2)

Input : $P^{\text{fixed}}_{\hat{Y}_1|Y_1}$, $P^{\text{init}}_{\hat{Y}_2|Y_2}$, $P_{Y_1,Y_2|X_1,X_2}$, $|\hat{\mathcal{Y}}_2|$, β_2, ϵ_2

Output : $[P^{\text{optimal}}_{\hat{Y}_2|Y_2}, c_2, R_{\text{sum}}]$

begin

Initialization: Set $t \leftarrow 0$, then set the initial mapping $P^{(0)}_{\hat{Y}_2|Y_2} \leftarrow P^{\text{init}}_{\hat{Y}_2|Y_2}$.

do

Based on $P^{\text{fixed}}_{\hat{Y}_1|Y_1}$ and newly obtained $P^{(t)}_{\hat{Y}_2|\hat{Y}_2}$, compute and update $d^{(t)}(y_2,\hat{y}_2) \leftarrow \beta_2 \sum_{\hat{y}_1} P_{\hat{Y}_1|Y_2} D_{\text{KL}}\left(P_{X_1X_2|\hat{Y}_1Y_2} || P^{(t)}_{X_1X_2|\hat{Y}_1\hat{Y}_2}\right) - \sum_{\hat{y}_1} P_{\hat{Y}_1|Y_2} \log_2\left(P^{(t)}_{\hat{Y}_2|\hat{Y}_1}\right) + \log_2\left(P^{(t)}_{\hat{Y}_2}\right)$.

Set $P^{(t+1)}_{\hat{Y}_2|Y_2} \leftarrow P^{(t)}_{\hat{Y}_2} 2^{-d^{(t)}(y_2,\hat{y}_2)} / \sum_{\hat{y}_2} P^{(t)}_{\hat{Y}_2} 2^{-d^{(t)}(y_2,\hat{y}_2)}$.

Set $t \leftarrow t+1$.

while $\sum_{y_2,\hat{y}_2} \left| P^{(t)}_{\hat{Y}_2|Y_2} - P^{(t-1)}_{\hat{Y}_2|Y_2} \right| / (|\mathcal{Y}_2| \cdot |\hat{\mathcal{Y}}_2|) \geq \epsilon_2$

Set $P^{\text{optimal}}_{\hat{Y}_2|Y_2} \leftarrow P^{(t)}_{\hat{Y}_2|Y_2}$, then compute $c_2 = I(Y_2;\hat{Y}_2|\hat{Y}_1)$ and $R_{\text{sum}} = I(X_1,X_2;\hat{Y}_1,\hat{Y}_2)$ based on it.

Function IB1($P^{\text{init}}_{\hat{Y}_1|Y_1}$, $P^{\text{fixed}}_{\hat{Y}_2|Y_2}$, $P_{Y_1,Y_2|X_1,X_2}$, $|\hat{\mathcal{Y}}_1|$, β_1, ϵ_1)

Input : $P^{\text{init}}_{\hat{Y}_1|Y_1}$, $P^{\text{fixed}}_{\hat{Y}_2|Y_2}$, $P_{Y_1,Y_2|X_1,X_2}$, $|\hat{\mathcal{Y}}_1|$, β_1, ϵ_1

Output : $[P^{\text{optimal}}_{\hat{Y}_1|Y_1}, c_1, R_{\text{sum}}]$

begin

Initialization: Set $t \leftarrow 0$, then set the initial mapping $P^{(0)}_{\hat{Y}_1|Y_1} \leftarrow P^{\text{init}}_{\hat{Y}_1|Y_1}$.

do

Based on $P^{\text{fixed}}_{\hat{Y}_2|Y_2}$ and newly obtained $P^{(t)}_{\hat{Y}_1|\hat{Y}_1}$, compute and update $d^{(t)}(y_1,\hat{y}_1) \leftarrow \beta_1 \sum_{\hat{y}_2} P_{\hat{Y}_2|Y_1} D_{\text{KL}}\left(P_{X_1X_2|Y_1\hat{Y}_2} || P^{(t)}_{X_1X_2|\hat{Y}_1\hat{Y}_2}\right)$.

Set $P^{(t+1)}_{\hat{Y}_1|Y_1} \leftarrow P^{(t)}_{\hat{Y}_1} 2^{-d^{(t)}(y_1,\hat{y}_1)} / \sum_{\hat{y}_1} P^{(t)}_{\hat{Y}_1} 2^{-d^{(t)}(y_1,\hat{y}_1)}$.

Set $t \leftarrow t+1$.

while $\sum_{y_1,\hat{y}_1} \left| P^{(t)}_{\hat{Y}_1|Y_1} - P^{(t-1)}_{\hat{Y}_1|Y_1} \right| / (|\mathcal{Y}_1| \cdot |\hat{\mathcal{Y}}_1|) \geq \epsilon_1$

Set $P^{\text{optimal}}_{\hat{Y}_1|Y_1} \leftarrow P^{(t)}_{\hat{Y}_1|Y_1}$, then compute $c_1 = I(Y_1;\hat{Y}_1)$ and $R_{\text{sum}} = I(X_1,X_2;\hat{Y}_1,\hat{Y}_2)$ based on it.

the noise level can be estimated by the BBU pool, where the optimization is performed. Hence, the probability distribution $P_{Y_1,Y_2|X_1,X_2}$, which describes the radio access channel between UEs and eRRHs is known and acts as an input parameter. Before the optimization starts, the cardinality of the compression index, i.e., $|\hat{\mathcal{Y}}_1|$ shall also be predetermined. It is usually determined by how much fronthaul capacity is available: When the fronthaul has more capacity, the cardinality can be set larger, which can support finer quantization and preserve more relevant information. Parameter ϵ_2 denotes a predetermined tolerance factor, which is used to determine when to terminate the algorithm.

$\beta_2 > 0$ is the input trade-off factor. As introduced in Subsection 2.2, when the conventional IB method is adopted, β is the input parameter of the method. Different values of $\beta > 0$ will yield different optimal points on the trade-off curve, via the IB method. Hence, different optimal trade-off points $\{I(Y_2;\hat{Y}_2|\hat{Y}_1), I(X_1,X_2;\hat{Y}_1,\hat{Y}_2)\}$ can be acquired, by inserting different values of β_2 to **Function IB2**. Steps 2 to 6 in **Function IB2** are the iterative optimization steps for $P_{\hat{Y}_2|Y_2}$ until reaching the convergence, which follows the instructions of the IB method [TPB99]. Based on the input probability distributions of **Function IB2**, all required probability distributions used in step 3 to calculate $d^{(t)}(y_2,\hat{y}_2)$, can be derived with basic rules introduced in the probability theory. Then in step 7, the optimized compression rate c_2, which is associated with the input value of β_2, and the corresponding maximized $I(X_1,X_2;\hat{Y}_1,\hat{Y}_2)$ are acquired. As the derivation of these iterative steps are the same as the conventional IB method, it is omitted here, mathematical details can be found in [TPB99] or [Zei11]. In Fig. 3.3, we fix a valid $P_{\hat{Y}_1|Y_1}$, by ranging β_2 from 0.1 to 50 and running proposed Function IB2 repeatedly, the concave trade-off curve in blue is plotted.

2. Similarly, when the second quantizer $P_{\hat{Y}_2|Y_2}$ is fixed and the chain rule is adopted, the trade-off between $\max\limits_{P_{\hat{Y}_1|Y_1}} I(X_1,X_2;\hat{Y}_1,\hat{Y}_2) = I(X_1,X_2;\hat{Y}_2) + \max\limits_{P_{\hat{Y}_1|Y_1}} I(X_1,X_2;\hat{Y}_1|\hat{Y}_2)$ and the compression rate $c_1 = I(Y_1;\hat{Y}_1)$ can also be obtained with the conventional IB method, as summarized in **Function IB1**. In this case, the fixed quantizer of eRRH 2, i.e., $P^{\text{fixed}}_{\hat{Y}_2|Y_2}$, becomes the input, and $\beta_1 > 0$ denotes the trade-off factor. In Fig. 3.3, the corresponding concave trade-off curve is plotted in red.

Remarks on Fig. 3.3: As introduced above, by running **Function IB1** and **IB2**, a specific value of β_1 corresponds only to a specific point on the red curve, and a specific value of β_2 will generate a specific point on the blue curve. In this figure, when we set $\beta_1 = \beta_2 = 0.1$ as the input parameter to **Function IB1** and **IB2**, the **leftmost** points on the red and blue curve can be obtained, respectively. When we increase their values, more optimal trade-off points are acquired towards right. We say they are optimal trade-off points, as with a specific compression rate (x-axis),

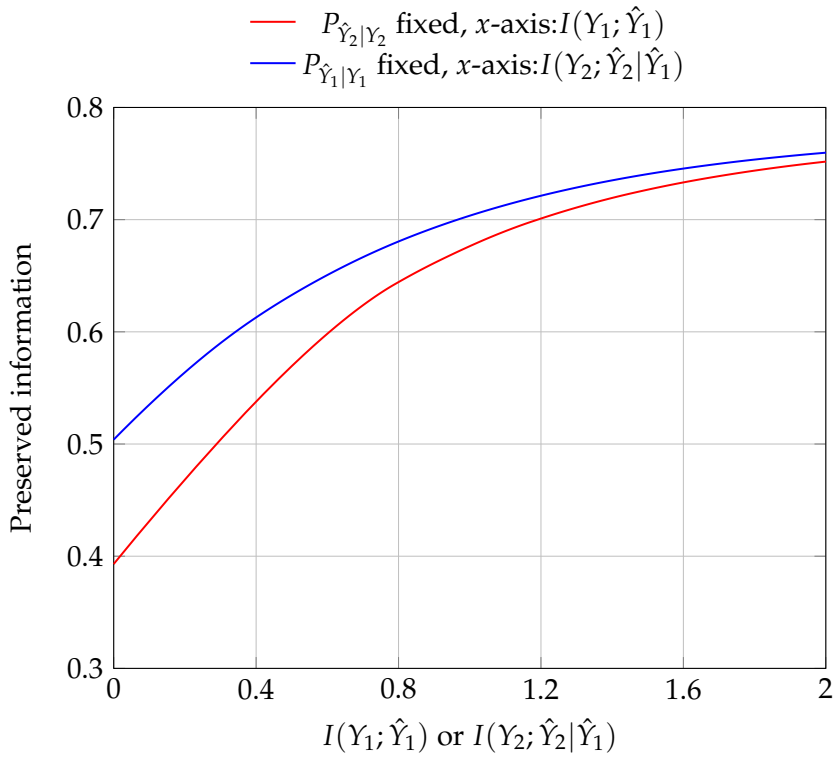


Figure 3.3: The trade-off between the preserved information $I(X_1, X_2; \hat{Y}_1, \hat{Y}_2)$ and the compression rates. BPSK modulation, $h_{11} = 1$, $h_{12} = 0.4$, $h_{21} = 0.6$, $h_{22} = 0.9$, $P_1 = 1$, $P_2 = 0.5$, $\sigma_n^2 = 1$, $|\hat{\mathcal{Y}}_1| = |\hat{\mathcal{Y}}_2| = 8$, $\epsilon_1 = 0.0003$, β_1, $\beta_2 \in [0.1, 50]$.

the curve depicts the **upper-bound** such that how much relevant information can be preserved. In the C-RAN/F-RAN model considered in this work, by increasing the value of β_1 and β_2, the fronthauls have to support the transmission of higher compression rate, but more relevant information can be finally preserved for the uplink transmission.

Then we go back to the original problem (3.2). As the signals received at two eRRHs are correlated and the Wyner-Ziv coding is executed for compression, the two quantizers shall be optimized jointly. In other words, the optimization of one quantizer always influences the optimization of the other, i.e., $I(\hat{Y}_1|Y_1)$ depends also on $P_{\hat{Y}_2|Y_2}$, and vice versa. We tackle this problem in an alternating manner: The trade-off between the correlated compression rate pair $\big(I(\hat{Y}_1;Y_1),\ I(\hat{Y}_2;Y_2|\hat{Y}_1)\big)$, and $I(X_1, X_2; \hat{Y}_1, \hat{Y}_2)$ can be obtained, by running **Function IB1** and **IB2** alternatively, such that the optimized quantizer obtained from one IB function is the input fixed quantizer of the other, until reaching the convergence. Such an Alternating Information Bottleneck (AIB) method is summarized in Function AIB.

Similarly to the conventional IB method, in the AIB method proposed for the two-eRRH case, a specific trade-off factor pair (β_1, β_2) corresponds to a specific compression rate pair of the two eRRHs. Furthermore, the corresponding upper-bound of the preserved information with this compression rate pair can be acquired via the

Function AIB($|\hat{\mathcal{Y}}_1|$, $|\hat{\mathcal{Y}}_2|$, $P_{Y_1,Y_2|X_1,X_2}$, β_1, β_2, ϵ_1, ϵ_2, ϵ_{AIB})

Input : $|\hat{\mathcal{Y}}_1|$, $|\hat{\mathcal{Y}}_2|$, $P_{Y_1,Y_2|X_1,X_2}$, β_1, β_2, ϵ_1, ϵ_2, ϵ_{AIB}
Output : $[c_1,\ c_2,\ R_{\text{sum}}]$
OptionalOutput: $[P^{\text{optimal}}_{\hat{Y}_1|Y_1},\ P^{\text{optimal}}_{\hat{Y}_2|Y_2}]$

1 **begin**
  Initialization : Randomly select a valid initial mappings $P^{(0)}_{\hat{Y}_1|Y_1}$ and $P^{(0)}_{\hat{Y}_2|Y_2}$, then set $t \leftarrow 0$.
2 **do**
3 Execute **Function IB1**: $P^{(t+1)}_{\hat{Y}_1|Y_1} = \text{IB1}(P^{(t)}_{\hat{Y}_1|Y_1},\ P^{(t)}_{\hat{Y}_2|Y_2},\ |\hat{\mathcal{Y}}_1|,\ \beta_1,\ \epsilon_1)$.
4 Execute **Function IB2**: $P^{(t+1)}_{\hat{Y}_2|Y_2} = \text{IB2}(P^{(t+1)}_{\hat{Y}_1|Y_1},\ P^{(t)}_{\hat{Y}_2|Y_2},\ |\hat{\mathcal{Y}}_2|,\ \beta_2,\ \epsilon_2)$.
5 Set $t \leftarrow t+1$.
6 **while** $\sum_{y_1,\hat{y}_1} \left| P^{(t)}_{\hat{Y}_1|Y_1} - P^{(t-1)}_{\hat{Y}_1|Y_1} \right| / (|\mathcal{Y}_1| \cdot |\hat{\mathcal{Y}}_1|) + \sum_{y_2,\hat{y}_2} \left| P^{(t)}_{\hat{Y}_2|Y_2} - P^{(t-1)}_{\hat{Y}_2|Y_2} \right| / (|\mathcal{Y}_2| \cdot |\hat{\mathcal{Y}}_2|) \geq \epsilon_{\text{AIB}}$
7 Set $P^{\text{optimal}}_{\hat{Y}_1|Y_1} \leftarrow P^{(t)}_{\hat{Y}_1|Y_1}$, $P^{\text{optimal}}_{\hat{Y}_2|Y_2} \leftarrow P^{(t)}_{\hat{Y}_2|Y_2}$, then compute $c_1 = I(Y_1;\hat{Y}_1)$, $c_2 = I(Y_2;\hat{Y}_2|\hat{Y}_1)$ and $R_{\text{sum}} = I(X_1,X_2;\hat{Y}_1,\hat{Y}_2)$ based on them.

AIB method. If the value of β_1 or β_2 is increased, higher corresponding compression rate at eRRH 1 or eRRH 2, and its associated compression strategy can be obtained with the AIB method, respectively. Moreover, the maximized preserved information can also be increased. Therefore, by inserting different value pairs of (β_1, β_2) into the AIB method, the optimal trade-off surface can be plotted. In step 7 of Function AIB, the optimized compression strategies for quantizers are acquired, which correspond to a specific point on the trade-off surface, i.e., the compression rate pair $\left(I(Y_1;\hat{Y}_1),\ I(Y_2;\hat{Y}_2|\hat{Y}_1)\right)$, that is associated with the trade-off factor pair (β_1, β_2) input to the algorithm, and the corresponding maximized preserved information $I(X_1, X_2; \hat{Y}_1, \hat{Y}_2)$. Such a trade-off surface will be given later in Fig. 3.6 of Subsection 3.5.1, where numerical results are provided. Before that, the convergence analysis of the AIB method shall be discussed.

3.2.2 Convergence Analysis

The proposed AIB method will definitely converge to at least a local optimal point. Note that in **Function IB2**, for the any fixed $P_{\hat{Y}_1|Y_1}$, it will definitely converge to the point where $I(X_1, X_2; \hat{Y}_2|\hat{Y}_1)$ is at least locally maximized, due to the convergence analysis of the conventional IB method [TPB99]. Therefore, the sum rate, or equivalently the preserved information of two UEs, i.e., $I(X_1, X_2; \hat{Y}_2, \hat{Y}_1) = I(X_1, X_2; \hat{Y}_1) + I(X_1, X_2; \hat{Y}_2|\hat{Y}_1)$ is also at least locally maximized, as the first term is temporarily fixed in the current alternating step. Then in the next step of the AIB method, the

optimized $P_{\hat{Y}_2|Y_2}$ is set as the fixed input parameter for **Function IB1**, with which $P_{\hat{Y}_1|Y_1}$ can be optimized, in order to further maximize $I(X_1, X_2; \hat{Y}_1|\hat{Y}_2)$, and thus $I(X_1, X_2; \hat{Y}_1, \hat{Y}_2)$. Hence, for specific compression rates, the value of $I(X_1, X_2; \hat{Y}_1, \hat{Y}_2)$ will at least **not decrease** in each alternating step, and thus can definitely converge to a local optimal point [RHL13]. As the problem is generally non-convex, similar to the convectional IB method, we can start with different initial points, i.e., different initial mappings $P^{(0)}_{\hat{Y}_1|Y_1}$ and $P^{(0)}_{\hat{Y}_2|Y_2}$, in the AIB method to acquire better results.

3.2.3 Extension to More UEs and eRRHs with Multiple Antennas

When more than two UEs and two eRRHs exist in the network, such an alternative mechanism can still be utilized, and as stated before, it is conceptually easy to be extended. In this scenario, for each eRRH, we still fix all the other quantizers, and optimize the quantizer of this specific eRRH with the conventional IB method. Then the optimized result is set as the fixed input, so as to optimize the next one. We can definitely do these steps alternatively until reaching the convergence, as the proof of the convergence for this case is similar to what we have introduced above. When multiple antennas are mounted on each eRRH, the received signals at different antennas of each eRRH are also correlated with each other, thus the Wyner-Ziv coding can still be utilized for the compression of them. In such a scenario, for any specific eRRH, the received signal at different antennas can be compressed by different quantizers. Each multiple-antenna eRRH can be regarded as being composed of several single-antenna *sub*-eRRHs. Hence, the proposed algorithm can still be adopted. But we must comment again that although such a conceptual extension is straightforward, the computational complexity increases exponentially.

3.3 The Alternating Bi-Section Method

Note that the AIB method can generate the optimal trade-off surface via inputting different values of the Lagrange multiplier vector $\boldsymbol{\beta}$ exhaustively. As we have stated several times above, each specific $\boldsymbol{\beta}$ corresponds to a specific compression rate vector, optimized compression strategies among all eRRHs, and the maximized preserved mutual information. For a specific network where the capacity of each fronthaul is know and fixed, it is better to know the exact value of vector $\boldsymbol{\beta}$, whose corresponding compression rate vector can exactly equal to the fronthaul capacity vector, in order to fully exploit the fronthaul capacity resources for preserving as much relevant information as possible. For locating vector $\boldsymbol{\beta}$ that exactly *matches* the available fronthaul capacity, we introduce a so-called alternating Bi-Section method, which is originated from the conventional Bi-Section method.

To be more specifically, we still take the scenario of two-UE two-eRRH as an example for simpler introduction. After setting up the trade-off surface numerically through the input trade-off pair (β_1, β_2) with the AIB method, we have to locate the point such that the constraints in (3.2) can be fulfilled simultaneously. Obviously, in order to fully exploit the fronthaul resource, the trade-off point where $I(Y_1; \hat{Y}_1) = C_{\text{FH},1}$ and $I(Y_2; \hat{Y}_2 | \hat{Y}_1) = C_{\text{FH},2}$ is to be found, then the corresponding maximized preserved information $I(X_1, X_2; \hat{Y}_1, \hat{Y}_2)$, as well as the compression strategies described by $P_{\hat{Y}_1|Y_1}$ and $P_{\hat{Y}_2|Y_2}$, are the solution for (3.2). Of course, we can achieve this by inserting different trade-off factor pairs (β_1, β_2), over a sufficiently fine grid of values in an exhaustive manner, until the point where $c_1 = C_{\text{FH},1}$ and $c_2 = C_{\text{FH},2}$ hold is finally found. Such an approach is apparently rather inefficient. As we have illustrated before, in the case of one quantizer, the compression rate c, which is resultant from the optimized compression strategy by the convectional IB method, increases by inputting larger value of β to the IB method [TPB99]. Thus, we can use the conventional Bi-Section method to locate any specific value of β, such that at the point on the trade-off curve where slope is $1/\beta$, the corresponding compression rate c exactly equals to the available capacity of the constraint link. For details please refer to [Zei11]. However, in C-RAN/F-RAN, there are multiple correlated quantizers, such that the resultant compression rate of a quantizer also depends on the achievable compression rates of the others. Therefore, the conventional *one-dimensional* Bi-Section method can not be directly utilized here. In order to deal with such a correlated compression scenario, we extend it to the alternating Bi-Section method below.

For a better illustration of the proposed algorithm, we execute the AIB function firstly, with different values of input trade-off factor pairs (β_1, β_2), and plot the resultant compression rate c_1 as the function of it, as depicted in Fig. 3.4. We see that the compression rate at eRRH 1, i.e., c_1, depends mainly on the input value of β_1: When β_2 is fixed, the value of c_1 is monotonically increasing with that of β_1, e.g., $(3, 1, 0.6581) - -(9, 1, 1.3357) - -(17, 1, 1.7892)$, which is the same as the conventional one quantizer scenario. Thus, we state that the value of c_1 is in *direct* association with the value of β_1. However, as the compression strategies between quantizers influence each other, the value of β_2 also slightly affects the value of c_1, as shown by the marked points in Fig. 3.4, e.g., $(9, 1, 1.3357) - -(9, 4, 1.2709) - -(9, 18, 1.1718)$. Thus, we state that the value of c_1 is in *indirect* association with that of β_2. If we adopt the conventional Bi-Section method to locate β_1 and β_2 individually, i.e., the value of β_1 is located where $c_1(\beta_1) = C_{\text{FH},1}$ fulfills, then we fix this β_1 and locate the value of β_2 until obtaining $c_2(\beta_2) = C_{\text{FH},2}$, the newly located value of β_2 (the indirect trade-off factor of c_1) will make c_1 slightly deviate from the previous value, and vice versa. Hence, the trade-off factor pair must be located somehow **jointly**, instead of independently with the conventional Bi-Section method. Similar to the AIB method, an alternating approach is proposed to achieve such a target: By

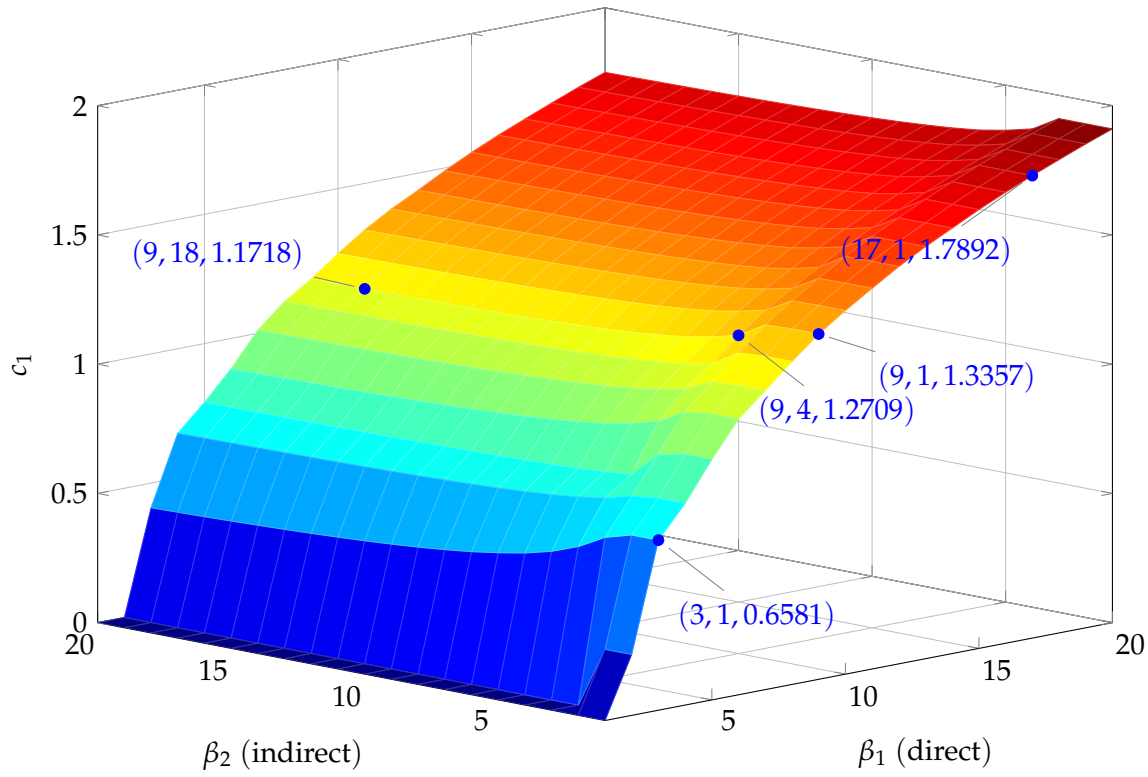


Figure 3.4: The relationship between the input trade-off factor pair (β_1, β_2) and the output compression rate c_1. The same channel setup of Fig. 3.3 is assumed.

fixing the value of the indirect associated trade-off factor β_j, $j \in \{1,2\}\backslash\{i\}$ for compression rate c_i, locating β_i and its directly associated compression rate c_i becomes the same as the one quantizer scenario, where the conventional Bi-Section method can be adopted. After the specific value of β_i is located, its associated compression rate now fulfills $c_i = C_{\mathrm{FH},i}$, under the condition that the other indirect trade-off factor is fixed. In the next alternating step, this newly located trade-off factor is fixed and the conventional Bi-Section method is adopted to locate the value of the other one. Such steps are executed alternatively until reaching convergence. The alternating Bi-Section method for the two-eRRH scenario is summarized in Alg. 1.

In the algorithm, $[\beta_{i\min}, \beta_{i\max}]$ indicates the searching range of β_i. $C_{\mathrm{FH},i}$ denotes the target compression rate for the i-th eRRH, which equals to its fronthaul capacity. η, ζ are the tolerance parameters used for terminating the Bi-Section search. From step 5 to step 13, the value of β_2 is fixed, and the Bi-Section method is executed to locate the value of β_1: At this point, its associated compression rate c_1 fulfills $c_1 = C_{\mathrm{FH},1}$. Then the value of β_2 is located, by fixing the value of β_1 from step 14 to step 22. These steps are executed repeatedly and alternatively until reaching convergence. After the trade-off factor pair is located via such an alternating manner, the AIB method is executed again in step 25, with which the corresponding optimal quantizers and the maximized sum rate are calculated. The optional output in Alg. 1 yields the compression rate pair (c_1, c_2), which is associated with the lo-

Algorithm 1: the alternating Bi-Section method

Input : $P_{Y_1,Y_2|X_1,X_2}$, $|\hat{\mathcal{Y}}_1|$, $|\hat{\mathcal{Y}}_2|$, $\beta_{1\max}$, $\beta_{1\min}$, $\beta_{2\max}$, $\beta_{2\min}$
Input : $C_{\text{FH},1}$, $C_{\text{FH},2}$, ϵ_1, ϵ_2, ϵ_{AIB}, η, ζ
Output : R_{sum}, $P^{\text{optimal}}_{\hat{Y}_1|Y_1}$, $P^{\text{optimal}}_{\hat{Y}_2|Y_2}$
OptionalOutput: c_1, c_2

begin
  Set $t \leftarrow 0$, $\beta_1^{(0)} \leftarrow (\beta_{1\max} + \beta_{1\min})/2$,
  $\beta_2^{(0)} \leftarrow (\beta_{2\max} + \beta_{2\min})/2$.
  do
    Set $\beta_{1\text{U}} \leftarrow \beta_{1\max}$, $\beta_{1\text{L}} \leftarrow \beta_{1\min}$
    while $\beta_{1\text{U}} - \beta_{1\text{L}} > \eta$ **do**
      Set $\widetilde{\beta}_1 \leftarrow (\beta_{1\text{U}} + \beta_{1\text{L}})/2$
      Execute Function AIB: $[c_1, \sim, \sim]$=AIB
       $(|\hat{\mathcal{Y}}_1|, |\hat{\mathcal{Y}}_2|, P_{Y_1,Y_2|X_1,X_2}, \widetilde{\beta}_1, \beta_2^{(t)}, \epsilon_1, \epsilon_2, \epsilon_{\text{AIB}})$
      if $c_1 < C_{\text{FH},1}$ **then**
        Set $\beta_{1\text{L}} \leftarrow \widetilde{\beta}_1$
      else
        Set $\beta_{1\text{U}} \leftarrow \widetilde{\beta}_1$
    Set $\beta_1^{(t+1)} \leftarrow (\beta_{1\text{U}} + \beta_{1\text{L}})/2$
    Set $\beta_{2\text{U}} \leftarrow \beta_{2\max}$, $\beta_{2\text{L}} \leftarrow \beta_{2\min}$
    while $\beta_{2\text{U}} - \beta_{2\text{L}} > \eta$ **do**
      Set $\widetilde{\beta}_2 \leftarrow (\beta_{2\text{U}} + \beta_{2\text{L}})/2$
      Execute Function AIB: $[\sim, c_2, \sim]$=AIB
       $(|\hat{\mathcal{Y}}_1|, |\hat{\mathcal{Y}}_2|, P_{Y_1,Y_2|X_1,X_2}, \beta_1^{(t+1)}, \widetilde{\beta}_2, \epsilon_1, \epsilon_2, \epsilon_{\text{AIB}})$
      if $c_2 < C_{\text{FH},2}$ **then**
        Set $\beta_{2\text{L}} \leftarrow \widetilde{\beta}_2$
      else
        Set $\beta_{2\text{U}} \leftarrow \widetilde{\beta}_2$
    Set $\beta_2^{(t+1)} \leftarrow (\beta_{2\text{U}} + \beta_{2\text{L}})/2$
    Set $t \leftarrow t + 1$
  while $|\beta_1^{(t)} - \beta_1^{(t-1)}| + |\beta_2^{(t)} - \beta_2^{(t-1)}| \geq \zeta$
  Execute Function AIB: $[c_1, c_2, R_{\text{sum}}, P^{\text{optimal}}_{\hat{Y}_1|Y_1}, P^{\text{optimal}}_{\hat{Y}_2|Y_2}]$=AIB
   $(|\hat{\mathcal{Y}}_1|, |\hat{\mathcal{Y}}_2|, P_{Y_1,Y_2|X_1,X_2}, \beta_1^{(t)}, \beta_2^{(t)}, \epsilon_1, \epsilon_2, \epsilon_{\text{AIB}})$

cated trade-off factor pair (β_1, β_2). We can easily verify whether such an algorithm generates correct results, by testing if $(c_1, c_2) = (C_{\text{FH},1}, C_{\text{FH},2})$ holds.

3.4 Fronthaul Capacity Allocation

As already stated in Section 1.2 when we introduce C-RAN, the fronthaul can be constructed via different technologies, e.g., the optical fiber communication, or the millimeter wave communication. The former one usually corresponds to the wired fronthaul, such that the capacity of each fronthaul is fixed and the infrastructures are deployed in advance. The fronthauling procedures to or from different eRRHs, do not have influences on each other. In the last subsection, where the quantizers are optimized via the proposed AIB method and the alternating Bi-Section method, such a network configuration is assumed: The fronthaul capacity for eRRH 1 and eRRH 2 are predetermined and fixed. Thus, we can locate value of (β_1, β_2), whose corresponding compression rate pair (c_1, c_2) can fully exploit the available fronthaul capacity $(C_{\text{FH},1}, C_{\text{FH},2})$. In contrast, the latter case usually corresponds to the wireless fronthaul, which is probably the only feasible realization, for the dense or heterogeneous 5G network [DC15]. In this case, it is usually not possible to deploy a large amount of fronthauls with, e.g., optical fibre, in advance. This is on one hand due to the high costs, it also might be, on the other hand, not efficient, as the traffic load on different fronthauls are not known in advance, and might even vary drastically over time. Therefore, a predetermined fronthaul capacity might lead to an inefficient operation of a practical network. Thanks to the wireless fronthaul, the fronthaul resources can be dynamically shared among all eRRHs. However, such non-dedicated fronthaul resources render the problem above much more complicated. As the capacity of each fronthaul is not predetermined, the allocation is also subject to optimization. Hence, no compression rate *target* is available to the proposed alternating Bi-Section method. Then the question is: How can we utilize the AIB method and the alternating Bi-Section method to locate any point, when it is even not known? Moreover, the fronthaul capacity allocation scheme to each eRRH apparently influences the design of the optimal compression strategies for quantizers, and different design of the quantizers also results in different resource allocation schemes. Such an interaction between the optimization of the compression strategies, and the fronthaul resource allocation scheme generates a more complicated problem, in which a joint optimization of both compression and resource allocation seems to be necessary, i.e., the BBU pool has to jointly optimize the compression procedure for each eRRH, as well as how much fronthaul capacity has to be allocated to it. This is the topic we are going to address in this subsection.

Furthermore, when it comes to the problem of the resource allocation, another issue is usually quite important: The QoS weights for different UEs: The QoS requirement

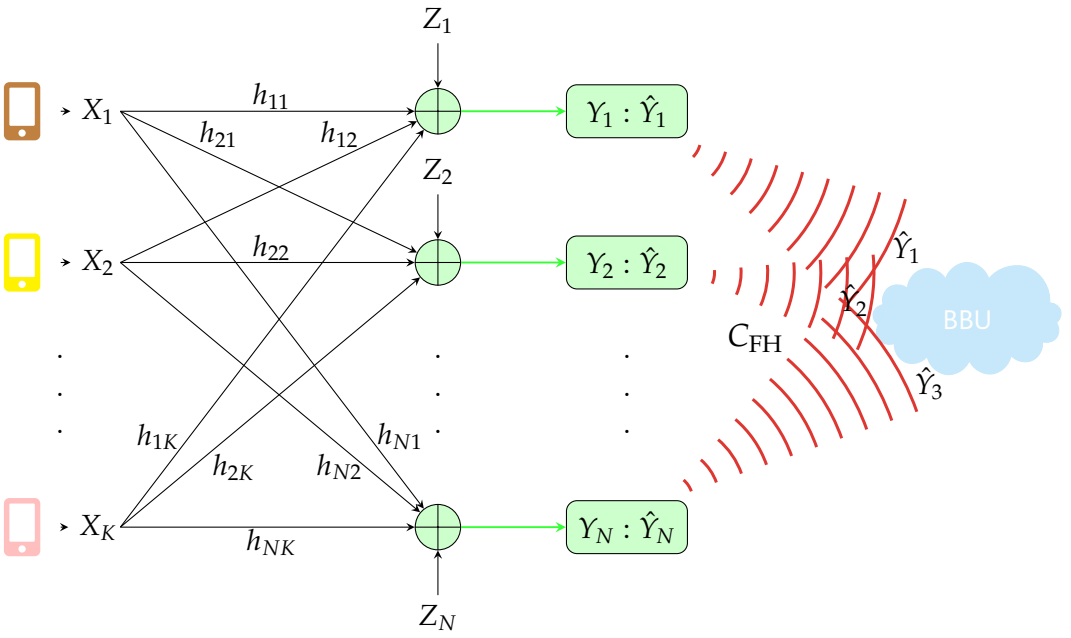


Figure 3.5: The abstract model for the uplink of F-RAN with non-dedicated fronthaul.

of a specific UE can be higher than that of the others. Hence, a finer compression strategy is expected for this UE, and the transmission of the information for this UE needs more biased fronthaul resource allocation. In this subsection, we further extend the proposed two methods, and utilize the Outer Linearization Method (OLM) [BSS16], to tackle such a complicated problem.

3.4.1 System Model and Problem Formulation

The system model we considered is illustrated in Fig. 3.5, where all fronthauls share the total capacity of C_{FH}. Different UEs have different predetermined QoS weights or priorities, according to the contents of their message. Such weights are assumed to be known at the BBU pool. We adopt w_k to denote the QoS weight of UE k. The larger the value of w_k is, the higher QoS requirement of this UE has. All other notations and assumptions are the same as in Subsection 3.1.7.

We aim to maximize the achievable **weighted** uplink sum rate [Par+14] for a F-RAN as follows:

$$\begin{aligned} &\max_{P_{\hat{Y}|Y}} \sum_{k=1}^{K} w_k R_k, \\ &\text{subject to } I(\boldsymbol{Y};\hat{\boldsymbol{Y}}) \leq C_{\text{FH}}, \end{aligned} \tag{3.3}$$

where $P_{\hat{\boldsymbol{Y}}|\boldsymbol{Y}} = \prod_{n=1}^{N} P_{\hat{Y}_n|Y_n}$. R_k denotes the achievable uplink rate of UE k. By manipulating the value of $\mathbf{w} = \{w_1, w_2, ..., w_K\}$, different QoS priorities can be granted to different UEs.

From the perspective of the BBU pool, the network acts as a MIMO-MAC, the capacity-achieving strategy in the MIMO-MAC is based on the well-known Successive Interference Cancellation (SIC) scheme [Gol12]. However, the optimal detection order for SIC is NP-hard. In practice, some predetermined fixed detection order is usually executed according to some criteria. According to [BW06], the solution of (3.3) is given by the detection order π that sorts the weights in a non-decreasing order

$$w_{\pi_1} \leq w_{\pi_2} \leq \cdots \leq w_{\pi_{k-1}} \leq w_{\pi_k}.$$

The UE with smaller QoS target is decoded before the UE with larger QoS weight coefficient, and all decoded symbols act as side information when the next symbols are decoded, so as to achieve higher rates for UEs with higher QoS targets. As in [BW06], such a decoding order has been shown to be close to optimal , we just adopt it here, as the the optimization for the SIC detection order is beyond the scope of this work.

Without loss of generality, we assume $w_K \geq w_{K-1} \geq \cdots \geq w_1$, i.e., symbol X_1 from UE 1 with the lowest QoS requirement is to be decoded at first, symbol X_K from UE K with the highest QoS requirement is to be decoded at last. Thus, the achievable rate for UE k can be written as

$$R_k = I(X_k; \hat{\mathbf{Y}} | X_1, X_2, ..., X_{k-1}),\ \forall k \in \{1, 2, ..., K\}. \tag{3.4}$$

The constraint of (3.3) can also be expressed as

$$\begin{aligned} &I(Y_n; \hat{Y}_n | \hat{Y}_1, \hat{Y}_2, ..., \hat{Y}_{n-1}) \leq C_{\mathrm{FH},n},\ \forall n \in \{1, 2, ..., N\}, \\ &\sum_{n=1}^{N} C_{\mathrm{FH},n} = C_{\mathrm{FH}}. \end{aligned} \tag{3.5}$$

where $C_{\mathrm{FH},n}$ denotes the fronthaul capacity allocated to eRRH n, which is subject to be optimized at the BBU pool. Obviously, a joint optimization of all compression strategies and the fronthaul capacity allocation is required. In order to make this problem more tractable and easier to be solved, we optimize them in a **sequential** and **iterative** way: Generally speaking, we suppose that the capacity of each fronthaul is predetermined at first, and optimize all quantizers jointly, via the proposed AIB method and the alternating Bi-Section method with the steps proposed in the previous subsections. Then we utilize the Outer Linearization Method (OLM), with which a mechanism is proposed, for the optimization of capacity allocation under the newly optimized compression strategies. Then in the next loop, the resultant fronthaul capacity allocation from OLM is regarded as known and predetermined, under which the compression strategies will be further updated.

3.4.2 Optimization with Predetermined Capacity Allocation

In this subsection, we will focus on the procedures to maximize the weighted uplink sum rate, by assuming that the fronthaul capacity allocation scheme is predetermined and known to the BBU pool. We will address the joint optimization of the resource allocation and the compression in the next subsection. Here, the proposed AIB method and the alternating Bi-Section method can be readily utilized to optimize the compression strategies, for UEs with different QoS requirements, under the current predetermined resource allocation scheme.

Similar to Subsection 3.2.1, for the ease of the introduction, we still assume that the F-RAN consists of two UEs and two eRRHs, and each is equipped with a single antenna. Moreover, as the resource allocation is supposed to be predetermined and known, the problem (3.3) can be expressed as follows:

$$\begin{aligned} \max_{P_{\hat{Y}_1|Y_1} P_{\hat{Y}_2|Y_2}} \quad & w_1 I(X_1;\hat{Y}_1,\hat{Y}_2) + w_2 I(X_2;\hat{Y}_1,\hat{Y}_2|X_1), \\ \text{subject to } & I(Y_1;\hat{Y}_1) \leq C_{\text{FH},1}, \\ & I(Y_2;\hat{Y}_2|\hat{Y}_1) \leq C_{\text{FH},2}. \end{aligned} \tag{3.6}$$

For the starting point, the resource allocation scheme, i.e., the value of $C_{\text{FH},1} > 0$ and $C_{\text{FH},2} > 0$ can be selected arbitrarily, as long as $C_{\text{FH},1} + C_{\text{FH},2} = C_{\text{FH}}$ is fulfilled. By expressing $R_{\text{wsum}} = w_1 I(X_1;\hat{Y}_1,\hat{Y}_2) + w_2 I(X_2;\hat{Y}_1,\hat{Y}_2|X_1)$ and adopting the chain rule and the proposed AIB method, we summarize the function to optimize the compression strategy for each eRRH as follows, which is quite similar to **Function IB1** and **IB2** in Subsection 3.2.1. Therefore, they are introduced here without derivations in detail, but the differences compared to **Function IB1** and **IB2** in Subsection 3.2.1 are highlighted in red.

After acquiring these two functions, we can adopt the corresponding AIB method to construct the optimal trade-off surface, and utilize the proposed alternating Bi-Section method to locate specific points on the trade-off surface. The procedure is the same as introduced in Subsection 3.2.1 and Section 3.3.

3.4.3 The Overall Algorithm for Fronthaul Capacity Allocation

In the last subsection, the fronthaul capacity allocated to each eRRH is supposed to be predetermined and known to the BBU pool. Then we adopted the AIB method and the alternating Bi-Section method to optimize the compression strategies, such that the corresponding compression rate vector can exactly fully exploit the predetermined fronthaul capacities simultaneously. In this subsection, we address the

Function wIB1($P^{\text{init}}_{\hat{Y}_1|Y_1}$, $P^{\text{fixed}}_{\hat{Y}_2|Y_2}$, $P_{Y_1,Y_2|X_1,X_2}$, $|\hat{\mathcal{Y}}_1|$, β_1, ϵ_1)

Input : $P^{\text{init}}_{\hat{Y}_1|Y_1}$, $P^{\text{fixed}}_{\hat{Y}_2|Y_2}$, $P_{Y_1,Y_2|X_1,X_2}$, $|\hat{\mathcal{Y}}_1|$, β_1, ϵ_1

Output : $[P^{\text{optimal}}_{\hat{Y}_1|Y_1}, c_1, R_{\text{wsum}}]$

begin

  Initialization: Set $t \leftarrow 0$, then set the initial mapping $P^{(t)}_{\hat{Y}_1|Y_1} \leftarrow P^{\text{init}}_{\hat{Y}_1|Y_1}$.

  do

    Based on $P^{\text{fixed}}_{\hat{Y}_2|Y_2}$ and newly obtained $P^{(t)}_{\hat{Y}_1|Y_1}$, compute and update

    $d^{(t)}(y_1,\hat{y}_1) \leftarrow w_1\beta_1 \sum_{\hat{y}_2} P_{\hat{Y}_2|Y_1} D_{\text{KL}}\left(P_{X_1X_2|Y_1\hat{Y}_2} || P^{(t)}_{X_1X_2|\hat{Y}_1\hat{Y}_2}\right) +$

    $(w_2 - w_1)\beta_1 \sum_{\hat{y}_2} P_{\hat{Y}_2X_2|Y_1} D_{\text{KL}}\left(P_{X_1|Y_1\hat{Y}_2X_2} || P^{(t)}_{X_1|\hat{Y}_1\hat{Y}_2X_2}\right)$.

    Set $P^{(t+1)}_{\hat{Y}_1|Y_1} \leftarrow P^{(t)}_{\hat{Y}_1} 2^{-d^{(t)}(y_1,\hat{y}_1)} / \sum_{\hat{y}_1} P^{(t)}_{\hat{Y}_1} 2^{-d^{(t)}(y_1,\hat{y}_1)}$.

    Set $t \leftarrow t+1$.

  while $\sum_{y_1,\hat{y}_1} \left| P^{(t)}_{\hat{Y}_1|Y_1} - P^{(t-1)}_{\hat{Y}_1|Y_1} \right| / (|\mathcal{Y}_1| \cdot |\hat{\mathcal{Y}}_1|) \geq \epsilon_1$

  Set $P^{\text{optimal}}_{\hat{Y}_1|Y_1} \leftarrow P^{(t)}_{\hat{Y}_1|Y_1}$, then compute $c_1 = I(Y_1;\hat{Y}_1)$ and R_{wsum} based on it.

Function wIB2($P^{\text{fixed}}_{\hat{Y}_1|Y_1}$, $P^{\text{init}}_{\hat{Y}_2|Y_2}$, $P_{Y_1,Y_2|X_1,X_2}$, $|\hat{\mathcal{Y}}_2|$, β_2, ϵ_2)

Input : $P^{\text{fixed}}_{\hat{Y}_1|Y_1}$, $P^{\text{init}}_{\hat{Y}_2|Y_2}$, $P_{Y_1,Y_2|X_1,X_2}$, $|\hat{\mathcal{Y}}_2|$, β_2, ϵ_2

Output : $[P^{\text{optimal}}_{\hat{Y}_2|Y_2}, c_2, R_{\text{wsum}}]$

begin

  Initialization: Set $t \leftarrow 0$, then set the initial mapping $P^{(t)}_{\hat{Y}_2|Y_2} \leftarrow P^{\text{init}}_{\hat{Y}_2|Y_2}$.

  do

    Based on $P^{\text{fixed}}_{\hat{Y}_1|Y_1}$ and newly obtained $P^{(t)}_{\hat{Y}_2|Y_2}$, compute and update

    $d^{(t)}(y_2,\hat{y}_2) \leftarrow$

    $w_1\beta_2 \sum_{\hat{y}_1} P_{\hat{Y}_1|Y_2} D_{\text{KL}}\left(P_{X_1X_2|\hat{Y}_1Y_2} || P^{(t)}_{X_1X_2|\hat{Y}_1\hat{Y}_2}\right) - \sum_{\hat{y}_1} P_{\hat{Y}_1|Y_2} \log_2\left(P^{(t)}_{\hat{Y}_2|\hat{Y}_1}\right) +$

    $\log_2\left(P^{(t)}_{\hat{Y}_2}\right) + (w_2 - w_1)\beta_2 \sum_{\hat{y}_1} P_{\hat{Y}_1X_2|Y_2} D_{\text{KL}}\left(P_{X_1|\hat{Y}_1Y_2X_2} || P^{(t)}_{X_1|\hat{Y}_1\hat{Y}_2X_2}\right)$.

    Set $P^{(t+1)}_{\hat{Y}_2|Y_2} \leftarrow P^{(t)}_{\hat{Y}_2} 2^{-d^{(t)}(y_2,\hat{y}_2)} / \sum_{\hat{y}_2} P^{(t)}_{\hat{Y}_2} 2^{-d^{(t)}(y_2,\hat{y}_2)}$.

    Set $t \leftarrow t+1$.

  while $\sum_{y_2,\hat{y}_2} \left| P^{(t)}_{\hat{Y}_2|Y_2} - P^{(t-1)}_{\hat{Y}_2|Y_2} \right| / (|\mathcal{Y}_2| \cdot |\hat{\mathcal{Y}}_2|) \geq \epsilon_2$

  Set $P^{\text{optimal}}_{\hat{Y}_2|Y_2} \leftarrow P^{(t)}_{\hat{Y}_2|Y_2}$, then compute $c_2 = I(Y_2;\hat{Y}_2|\hat{Y}_1)$ and R_{wsum} based on it.

optimization of the fronthaul capacity allocation, where the total capacity is to be allocated among eRRHs, so as to maximize the weighted uplink sum rate of the F-RAN. The proposed mechanism combines the AIB method, the alternating Bi-Section method, and the Outer Linearization Method (OLM) [BSS16].

Note that in the original problem (3.3), the objective function is concave, with respect to the compression rates. Moreover, the sum of all compression rates is limited by the sum capacity C_{FH}, which is also a linear inequality constraint. Thus, the original problem (3.3) is actually a convex optimization problem, with respect to the eRRH compression rate vector $\boldsymbol{c}$. As the proposed AIB method and the alternating Bi-Section method can be utilized for calculating the value of the objective function, for different vectors $\boldsymbol{c}$, the original problem (3.3) can be solved by standard convex optimization methods in an iterative manner. Here, similar to [Zei11], we can utilize the OLM to achieve this target. The overall procedures are listed as follows:

1. Select an arbitrarily valid capacity allocation, $\boldsymbol{C}^{(0)} = \left[C_{\mathrm{FH},1}^{(0)}, C_{\mathrm{FH},2}^{(0)}, ..., C_{\mathrm{FH},N}^{(0)}\right]^T$, such that $\sum_{n=1}^{N} C_{\mathrm{FH},n}^{(0)} = C_{\mathrm{FH}}$ is fulfilled. Set $\ell = 0$, $f_{\mathrm{LB}} = -1$ and f_{UB} to be large enough. Set δ be the predetermined tolerance factor for terminating the algorithm.

   Repeat step 2 to step 4 as below until $f_{\mathrm{UB}} - f_{\mathrm{LB}} \leq \delta$**.**

2. Use the proposed AIB and the alternating Bi-Section method to update the trade-off factor vector $\boldsymbol{\beta}^{(\ell)} = \left(\beta_1^{(\ell)}, \beta_2^{(\ell)}, ..., \beta_N^{(\ell)}\right)$, which is associated with the current capacity allocation scheme $\boldsymbol{C}^{(\ell)}$.

3. From step 2, the corresponding maximized weighted uplink sum rate $R_{\mathrm{wsum}}^{(\ell)}$ can be acquired. Set $f_{\mathrm{LB}} = R_{\mathrm{wsum}}^{(\ell)}$ and the sub-gradient $\boldsymbol{g}^{(\ell)} = \left(1/\beta_1^{(\ell)}, 1/\beta_2^{(\ell)}, ..., 1/\beta_N^{(\ell)}\right)$ and $b^{(\ell)} = R_{\mathrm{wsum}}^{(\ell)} - \boldsymbol{C}^{(\ell)} \cdot (\boldsymbol{g}^{(\ell)})^T$.

4. Then construct and solve the linear problem below

$$\max_{\boldsymbol{C}} s, \tag{3.7}$$

$$\text{s.t. } \boldsymbol{C} \cdot (\boldsymbol{g}^{(l)})^T + b^{(l)} \geq s,\ l = 0, 1, ..., \ell - 1, \tag{3.8}$$

$$\sum_{n=1}^{N} C_{\mathrm{FH},n} = C_{\mathrm{FH}}. \tag{3.9}$$

   Let $(s^*, \boldsymbol{C}^*)$ be the maximizer, set $f_{\mathrm{UB}} = s^*$, $\boldsymbol{C}^{(\ell+1)} = \boldsymbol{C}^*$, and $\ell = \ell + 1$.

In step 2 and 3 of the algorithm above, the optimized quantizers for a specific fronthaul capacity allocation scheme is obtained. Then in step 4, we fix the quantization

scheme but optimize only the fronthaul capacity allocation. Note that when the quantizers are all fixed, the original problem (3.3) is just a Linear Programming (LP) problem with respect to $\boldsymbol{C} = [C_{\mathrm{FH},1}, C_{\mathrm{FH},2}, ..., C_{\mathrm{FH},N}]^T$. Moreover, we can easily acquire the sub-gradient by inverting $\boldsymbol{\beta}$. The second constraint (3.9) guarantees that the newly generated allocation scheme fulfills the total capacity constraint. The constraints (3.8), together with the objective (3.7), aim to re-distribute the fronthaul capacity to reach a point, such that higher R_{wsum} can be achieved.

3.5 Numerical Results

In order to evaluate the performance of the proposed algorithms, in this section, some numerical results will be given. The environment is set up and simulated using MATLAB. For solving the optimization problems, we adopt CVX.

3.5.1 The AIB Method and the Alternating Bi-Section Method

In this subsection the performance of the proposed AIB method and the alternating Bi-Section method are to be evaluated, with which we are able to investigate its performance for the uplink transmission of the F-RAN, and derive some design guidelines.

General Setup: A network consisting of two-UE and two-eRRH is considered, each device is equipped with a single antenna. BPSK modulation is adopted for simplicity. The received analogue signal at each eRRH is sampled and discretized with 7 bits/sample, thus we have $|\mathcal{Y}_i| = 128$ before the compression executed by the quantizers. Then each eRRH will quantize the signals into eight quantization levels. i.e., at most 3 bits/sample, and then we have $|\hat{\mathcal{Y}}_i| = 8$. Moreover, we set $\beta_{1\max} = \beta_{2\max} = 260$, $\beta_{1\min} = \beta_{2\min} = 0.1$, $\epsilon_1 = \epsilon_2 = 3 \times 10^{-4}$, $\epsilon_{\mathrm{AIB}} = 10^{-5}$, $\eta = \zeta = 0.01$, which are the predetermined parameters required in the proposed algorithms.

At first the effectiveness and correctness of the proposed AIB method is to be verified, as shown in Table 3.1. We input different trade-off factor pairs (β_1, β_2) to the AIB function (in column 2), then the output are the compression rate pairs (c_1, c_2) (in column 3), which are associated with the input trade-off factor pairs (β_1, β_2), as well as the corresponding maximized uplink sum rates R_{sum} (in column 4). In other words, the data in column 3 and column 4 are the points on the trade-off surface, which is calculated via the proposed AIB method by inserting the factor pairs listed in column 1. Then we set (c_1, c_2), which is generated by the AIB method, as the input target rate pair to the alternating Bi-Section method, and utilize it to locate

Table 3.1: The comparison between the located points with the original ones, with $h_{11} = 1,\ h_{12} = 0.4,\ h_{21} = 0.6,\ h_{22} = 0.9,\ P_1 = 1,\ P_2 = 0.5$ and $\sigma_n^2 = 1$.

No.	(β_1, β_2)	(c_1, c_2)	R_{sum}	$(c_1^\star, c_2^\star)$	$R_{\text{sum}}^\star$
1	$(5, 8)$	$(0.7886, 0.9386)$	0.5599	$(0.7887, 0.9389)$	0.5600
2	$(25, 15)$	$(1.9244, 1.3326)$	0.7107	$(1.9243, 1.3328)$	0.7107
3	$(50, 50)$	$(2.3920, 2.1718)$	0.7578	$(2.3920, 2.1718)$	0.7578
4	$(10, 250)$	$(1.2306, 2.7249)$	0.7057	$(1.2305, 2.7249)$	0.7057
5	$(230, 20)$	$(2.8843, 1.5341)$	0.7424	$(2.8843, 1.5342)$	0.7424
6	$(260, 260)$	$(2.9003, 2.7049)$	0.7705	$(2.9003, 2.7049)$	0.7705

them. The AIB method can be claimed to work as expected and generate the correct results, as long as the output of the alternating Bi-Section method, i.e., $(c_1^\star, c_2^\star)$ (in column 5) and $R_{\text{sum}}^\star$ (in column 6) equals to the compression rate targets (c_1, c_2) and R_{sum}, respectively. From the results in Table 3.1, we observe that the algorithm has the capability, to locate the target compression rate pairs with high precision. Therefore, if the F-RAN adopts these algorithms to design the compression strategies, the resultant compression rates are able to meet the fronthaul capacities exactly to fully exploit the available fronthaul resources.

In the next step, we set different target compression rate pairs (c_1, c_2), and utilize the proposed AIB method to acquire the optimal trade-off surface, between the compression rate pair $\left(I(Y_1; \hat{Y}_1), I(Y_2; \hat{Y}_2|\hat{Y}_1)\right)$ and the corresponding maximized uplink sum rate $I(X_1, X_2; \hat{Y}_1, \hat{Y}_2)$, as shown in Fig 3.6. From the figure we can easily observe that it is a convex and increasing surface with respect to the compression rate pair, which is in line with the theory. The maximal uplink sum rate can be increased, if either compression rate is increased. In other words, if the fronthaul has higher capacity, the quantization steps can be finer, more relevant information can be preserved at the destination, and thus the uplink sum rate will be more and more close to the theoretical limit $I(X_1, X_2; Y_1, Y_2) = 0.9203$ (when no quantization needed, i.e., the fronthauls can support the transmission of the un-quantized signals) of this case. In particular, if the compression rate pair $(2.9, 2.7)$ can be supported by the fronthauls of eRRH 1 and eRRH 2, respectively, the total achievable rate will reach 0.9135, which we have marked on the trade-off surface.

As stated in Subsection 3.4.1, at the BBU pool, the decompression order of the compressed signals from different eRRHs can generally affect the achievable performance. We adopt the strategy such that the signals coming from the eRRHs with larger fronthaul capacity are decompressed at first, and then those with smaller ones. Now, the proposed algorithms are to be used for investigating the relationship between the decompression orders and the fronthaul capacities, as well as the reliability of the signals received by eRRHs. Hence, we manually allocate the total fronthaul resources to different eRRHs, in order to know the performance of the same decompression order for different capacity allocation schemes. Moreover, dif-

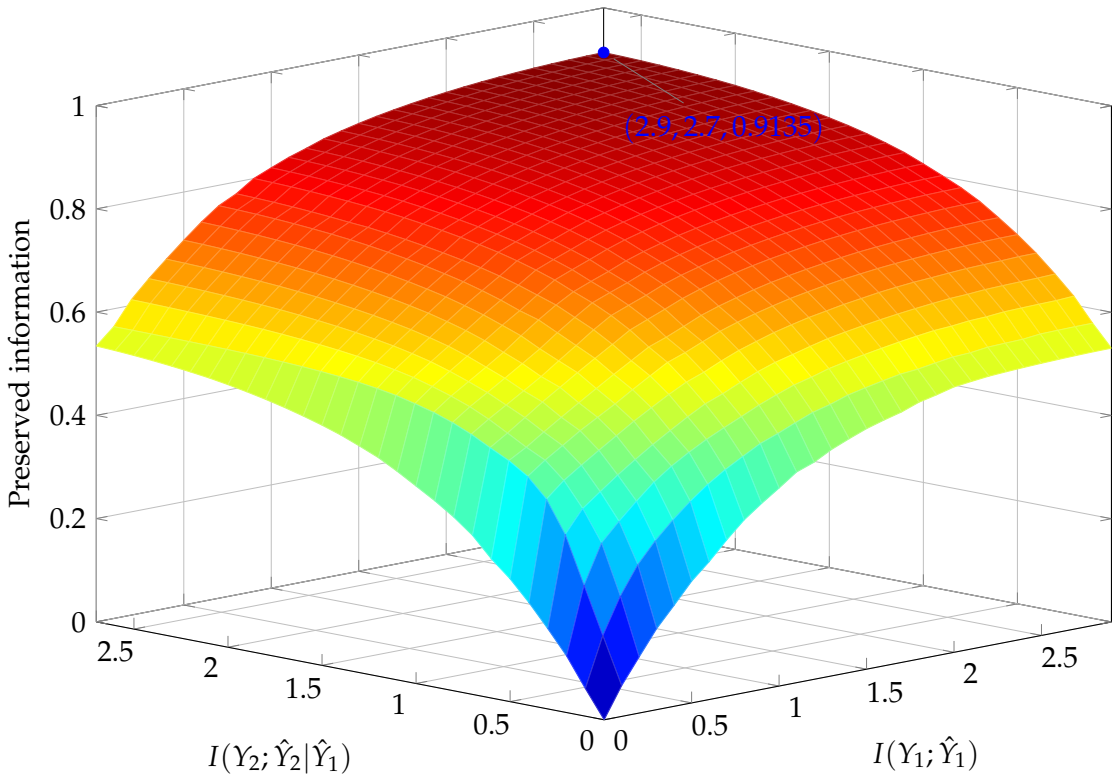


Figure 3.6: The trade-off surface between the compression rates and the preserved information $I(X_1, X_2; \hat{Y}_1, \hat{Y}_2)$, with $h_{11} = 1$, $h_{12} = 0.4$, $h_{21} = 0.6$, $h_{22} = 0.9$, $P_1 = P_2 = 1$, $w_1 = w_2 = 1$.

ferent noise levels are set at different eRRHs, i.e., $\sigma^2_{n,i}$ denotes the noise power at eRRH i. According to different assignments of noise power, we group the simulations into four cases, as shown in Fig. 3.7. The four groups are classified by different noise power and different SNR regimes at different eRRHs:

- **Group 1**: eRRH 1 and eRRH 2 experience the same noise level of 0.5, but different decompression orders are executed at the BBU pool (e.g., $1 \rightarrow 2$ denotes that the compressed signal from eRRH 1 is decompressed at first).
- **Group 2**: eRRH 1 and eRRH 2 experience different noise levels, the signal received by eRRH 2 is more reliable, as its noise power is lower.
- **Group 3**: Similar to group 1, but with higher noise levels at both eRRHs. Hence, the network works in lower SNR regime.
- **Group 4**: Similar to group 2, but the signal received by eRRH 1 is more reliable. Moreover, both eRRHs experience higher noise levels. Hence, the network works in lower SNR regime.

Furthermore, we use $\theta \in [0, 1]$ to denote the proportion of the total fronthaul capacity, that are manually allocated to the first eRRH, before the network start to operate.

Then the proposed algorithms are adopted to compute the maximal achievable sum rate in each case, with which the best fronthaul capacity allocation scheme, under a certain decompression order, can be derived. In Fig. 3.7, the marked points denote the optimal capacity allocations where the sum rate is maximized. By comparing the results of group 2, 3 and 4, we see that if the observation at one eRRH is more reliable than the other, allocating more fronthaul resources to this eRRH is preferred, so as to achieve higher uplink sum rate. Moreover, by comparing group 1 and 3, it can be concluded that when the capacity allocation factor θ is fixed somewhere already, and the reliability of the observations at two eRRHs are more or less the same, decompressing the signal from the eRRH with larger fronthaul capacity at first can yield better performance: When $\theta < 0.5$, i.e., more fronthaul capacity is allocated to eRRH 2, the decoding order $2 \rightarrow 1$ is better as a higher rate can be achieved than the other way around. When $\theta > 0.5$, i.e., more fronthaul capacity is allocated to eRRH 1, the results in Fig. 3.7 then demonstrate that the decoding order $1 \rightarrow 2$ shall be preferred in this case. Moreover, we observe that the performance gap between these two decompression orders becomes more pronounced in higher SNR regime, i.e., when the noise power σ^2 is smaller. This is due to the fact that, the useful signal becomes more and more dominant over the additional noise in this regime, thus the decompression order imposes more impact on the overall performance. Furthermore, we can also conclude that if the BBU pool has the flexibility to allocate capacity, i.e., it can manipulate the value of θ to maximize the uplink sum rate, the maximal achievable rates of these two orders are nearly the same. In the next subsection, the numerical results when the BBU pool has the ability to optimize the fronthaul capacity allocation will be given, instead of the manual allocation here.

Next, the performance of the Wyner-Ziv (WZ) coding with that of the Single Unit (SU) compression are compared. The Single Unit compression indicates that the compression strategy performed by each quantizer, ignores the correlation between the signals of the neighboring eRRHs. The BBU pool optimizes the quantizer of each eRRH in parallel, individually and independently: It only aims to maximize its own relevant information retrieval, without the consideration of exploiting the correlated information from other eRRHs. For example, for the n-th eRRH, the quantizer is designed such that $I(X_1, X_2; \hat{Y}_n)$ can be maximized with the compression rate $I(Y_n; \hat{Y}_n)$. Due to the individual and independent optimization steps in this scenario, the conventional IB method and the Bi-Section method can be directly applied. When the Wyner-Ziv coding is performed, the proposed AIB method for jointly optimizing the compression strategies is to be adopted. In the simulation, we change the available sum capacity and try different allocation factors θ until finding the optimal one for both cases. Moreover, different available power levels of the UE are also considered. The results are plotted in Fig. 3.8.

From this figure, we see that the WZ approach can uniformly outperform the SU ap-

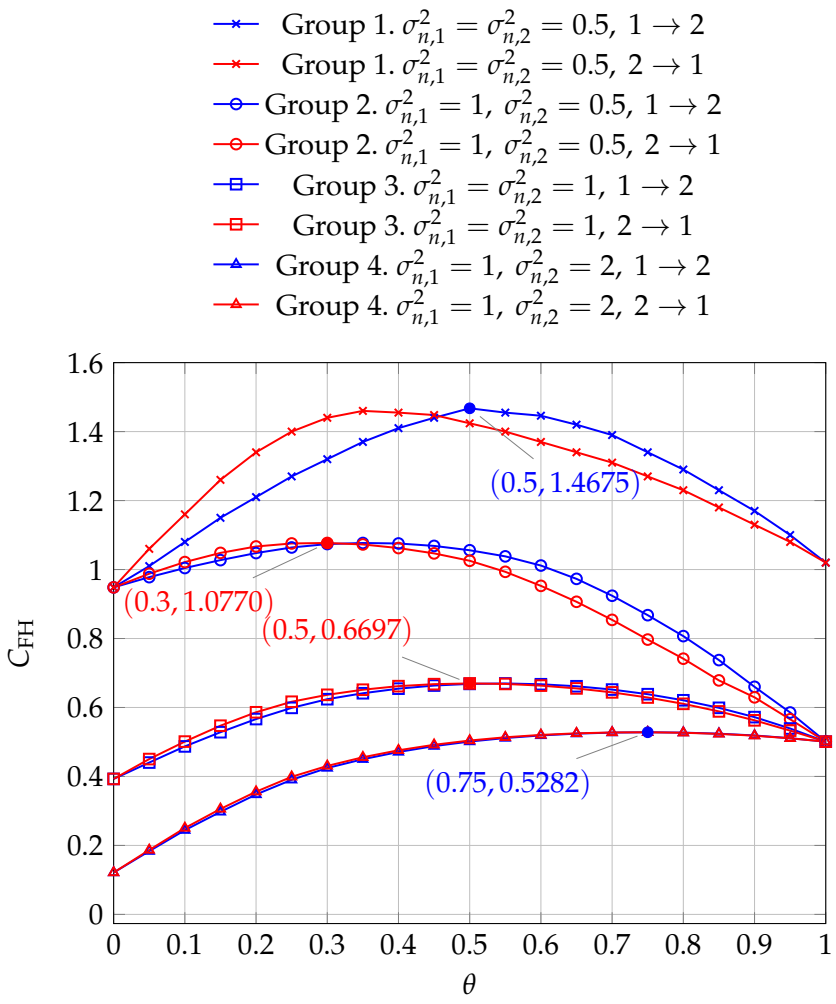
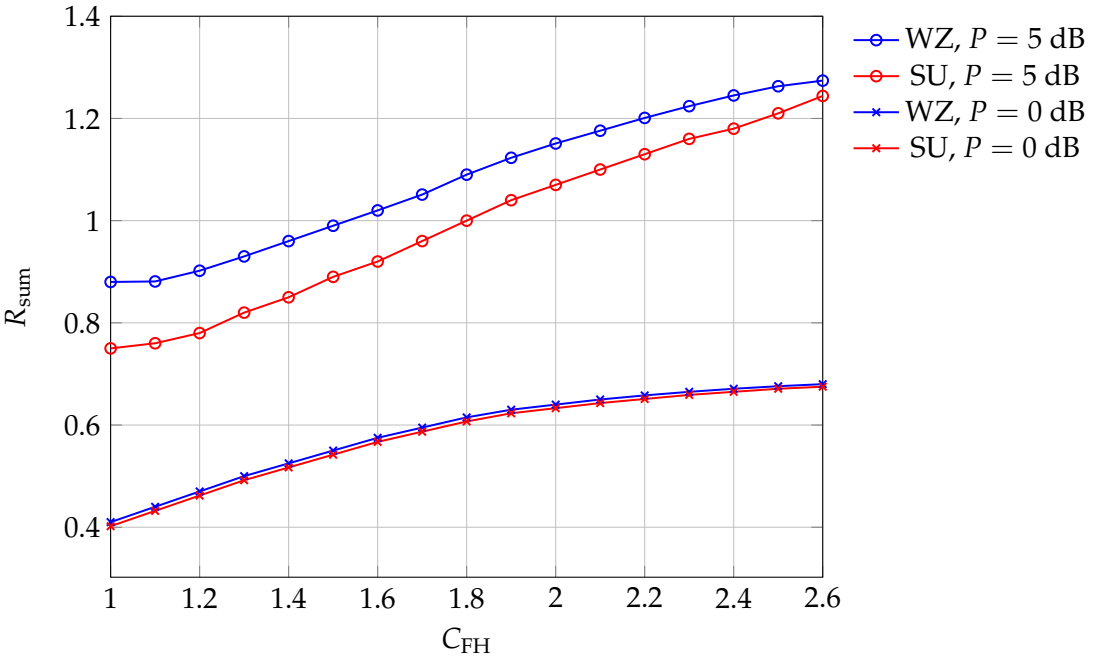


Figure 3.7: The relationship between the capacity allocation, decompression order and the maximal achievable sum rate. $i \to j$ denotes the signal from eRRH i is decompressed before that from eRRH j.

Figure 3.8: The comparison between the Wyner-Ziv coding with joint optimization, and the Single Unit compression, for different power levels of UE, and different fronthaul capacities.



proach, but at the cost of higher computational complexity: 1. The WZ approach has higher complexity from the aspect of implementation, as in the compression step, the correlations between the received signals from eRRHs are taken into consideration; 2. The optimization of the compression processes must be executed jointly in a centralized way, with the proposed AIB method and the alternating Bi-Section method. From the figure, it can be observed that the advantage of the Wyner-Ziv coding with joint optimization over the Single Unit compression becomes more apparent, as the available power levels of the UE get higher, which is in consistence with the theoretical analysis in [Sha14]. This is due to the fact that, the correlations between the received signals at neighboring eRRHs become more pronounced, when the network works in high SNR regimes. Moreover, when the sum capacity of the fronthaul becomes larger, such an advantage will be less prominent, which is in consistence with the theoretical analysis in [ZY14]. The reason is that in this scenario, the fronthaul capacity is not the main performance bottleneck anymore, sufficient available capacity exists already for the transmission of the compressed information, i.e., making better use of the fronthaul resources is not that urgent.

3.5.2 Fronthaul Capacity Allocation

In this subsection, we provide the numerical results when the BBU pool also has the freedom for the fronthaul resource allocation. The general set up of the network is similar to that of the last subsection, but we consider a F-RAN consisting of three

UEs and three eRRHs, each UE is with different QoS weights for their uploaded data streams. We set $w_3 = 3, w_2 = w_1 = 1$, i.e., the transmission of the data stream from the third UE is more prioritized than that of the other two. Moreover, the radio access channel is configured as $h_{11} = 1$, $h_{12} = 0.3$, $h_{13} = 0.2$, $h_{21} = 0.2$, $h_{22} = 1$, $h_{23} = 0.3$, $h_{31} = 0.2$, $h_{32} = 0.1$, $h_{33} = 0.5$, $\sigma^2_{n,1} = \sigma^2_{n,2} = \sigma^2_{n,3} = 1$. The decompression order is set as $1 \to 2 \to 3$.

At first the results for three different cases are compared:

- Case 1: The quantizers as well as the fronthaul capacity allocation are jointly optimized with the proposed AIB method, in order to maximize the sum rate without weighting, i.e., the priority of UE 3 is not considered.
- Case 2: The capacity allocation obtained from the results of Case 1 is adopted, the AIB method and the alternating Bi-Section method are adopted to optimize the quantizers only, so as to maximize the weighted sum rate, i.e., the joint optimization for compression and the fronthaul capacity allocation is not considered.
- Case 3: Both the quantizers and the capacity allocation are optimized jointly for maximizing the weighted sum rate. Hence, both issues, i.e., the fronthaul capacity allocation, and different significance of the data streams are taken into account.

By executing the proposed algorithms with the different configuration cases listed above, the achievable uplink rates for different UEs are documented, as well as the sum rate, with respect to the sum capacity of the fronthaul. The results are shown in Fig. 3.9 - 3.12.

From the figures we can easily observe that when the quantizers and the capacity allocation are optimized in order to maximize the sum rate, without considering the QoS weights, i.e., Case 1, the individual achievable rate of the third UE R_3 is the smallest, although it should have the most significance. If the QoS weight $w_3 = 3$ for R_3 is considered, but only the quantizers are to be optimized accordingly but without the resource allocation, i.e., Case 2, we see that it is not sufficient as shown in Fig. 3.11: The improvement of R_3 in Case 2 compared to the Case 1 is not significant. This is because the received signals at different eRRHs are the superposition of the signals from all UEs, only optimizing the compression strategies can not impose a significant impact on the individual achievable rates. In order to further improve the individual rate with larger QoS requirement, it is necessary to consider a simultaneous optimization of both fronthaul capacity allocation and the compression. From Fig. 3.11, we observe that the improvement of R_3 in Case 3 is much more

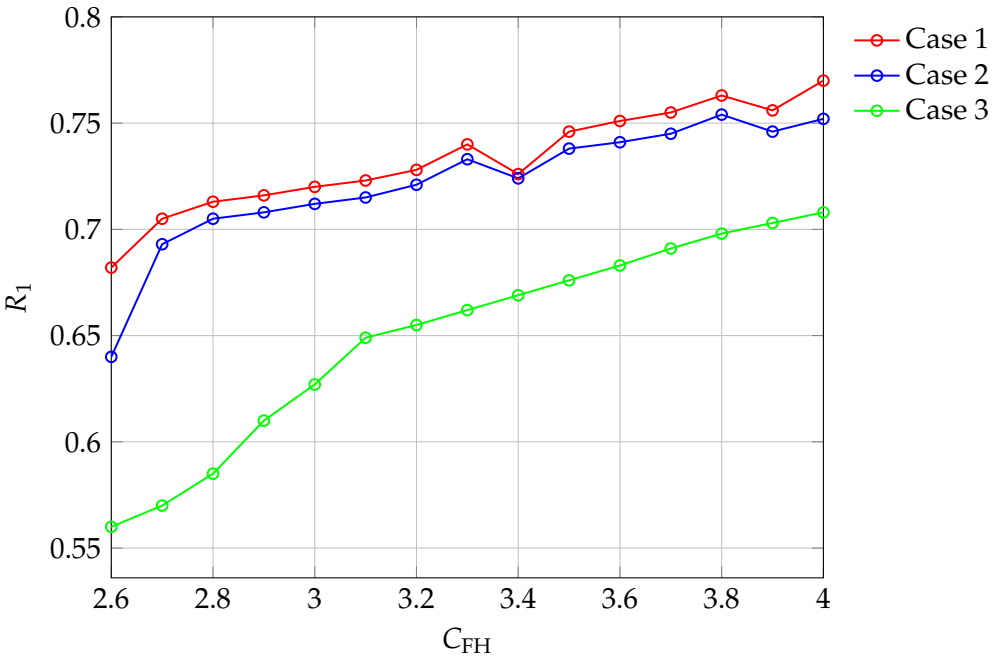


Figure 3.9: R_1 with respect to different sum capacities of the fronthaul.

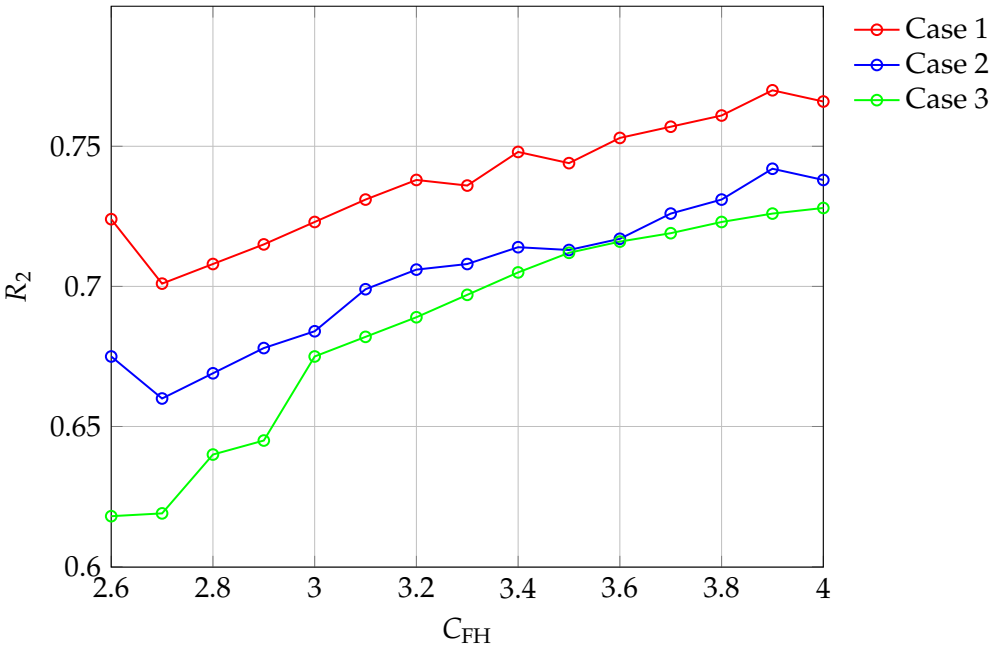


Figure 3.10: R_2 with respect to different sum capacities of the fronthaul.

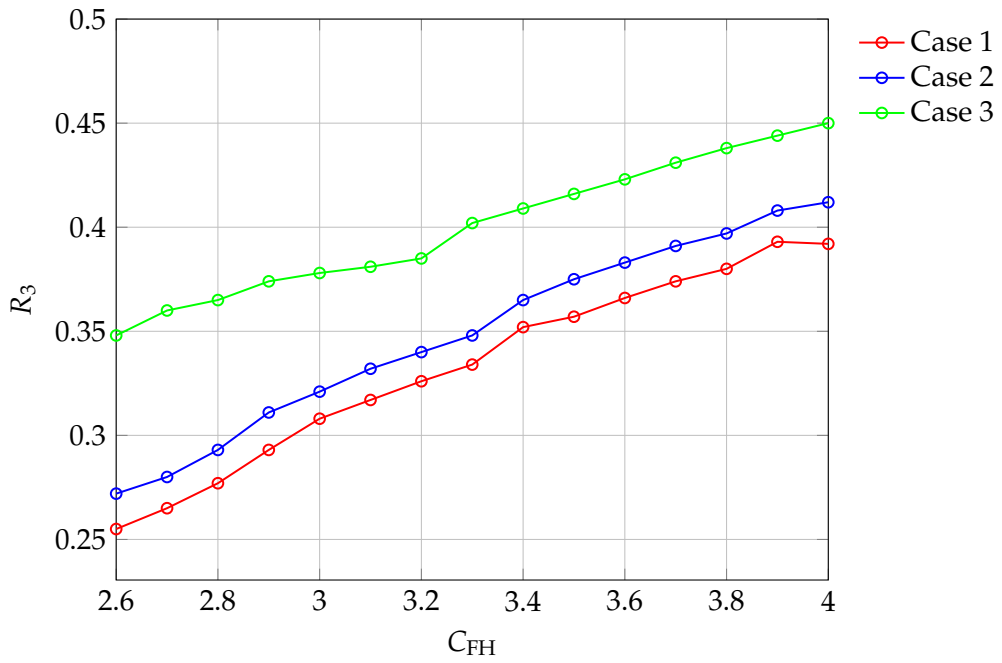


Figure 3.11: R_3 with respect to different sum capacities of the fronthaul.

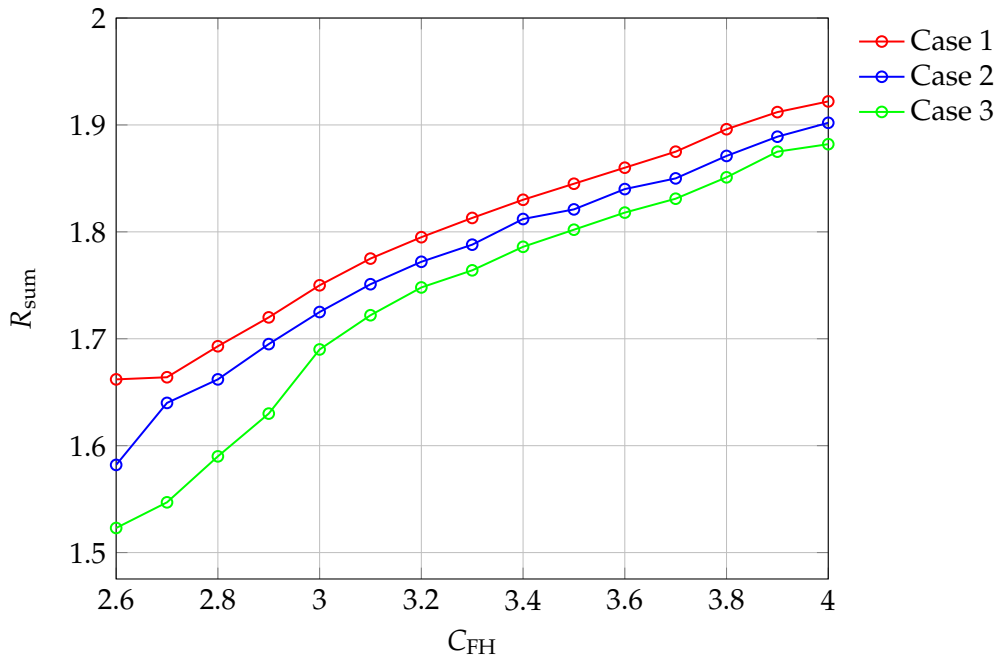


Figure 3.12: The sum rate with respect to different sum capacities of fronthaul.

prominent than that of Case 2. However, by comparing the results of other figures, i.e., the value of R_1, R_2 and R_{sum} for three cases, it can be concluded that such an improvement is at the cost of larger performance degradation of UE 1 and UE 2, as well as the sum rate R_{sum}. This is due to the fact that, more network resources are biased to support the data transmission for UE 3. Although it is beneficial to improve the QoS of UE 3, it leads to negative impacts on the overall performance of the network, as well as on other UEs.

Finally, by considering the same model as above, the optimal fronthaul capacity allocations obtained from the proposed algorithm for different optimization objectives, i.e., different QoS weights among UEs, is compared. The results are shown in Fig. 3.13 and Fig. 3.14. We see that when the compression strategies and the fronthaul capacity allocation are optimized for maximizing the uplink sum rate, only 18% of the capacity shall be allocated to eRRH 3. While if UE 3 is given more priority by maximizing the weighted sum rate (with $w_3 = 3$, $w_1 = w_2 = 1$), 38% of the capacity shall now be allocated to eRRH 3. For avoiding confusions, we emphasize here that eRRH 3 is not solely responsible for UE 3. Actually, each eRRH receives a superposed signals from all UEs, i.e., the signal received by each eRRH contains useful information of each UE. The only difference between eRRHs is that, some eRRHs might receive more powerful signal from a specific UE, e.g., this UE is more close to them. Hence, if more fronthaul capacity resources are allocated to these eRRHs, more information of this specific UE can be preserved finally at the BBU pool. Then we go back to the simulation results, the reason why more fronthaul capacity is allocated to eRRH 3, when UE 3 is considered to have higher priority, is that the signal from UE 3 at eRRH 3 is the strongest (note that we configure the channel gain as $h_{13} = 0.2$, $h_{23} = 0.3$, $h_{33} = 0.5$), and at eRRH 1 it is the weakest. Hence, the observation of the signal from UE 3 is most reliable at eRRH 3. Better compression strategy shall be considered at this eRRH if preserving more information from UE 3 is desired. For this specific F-RAN realization in the simulation, there are more fronthaul resources allocated to this eRRH, when UE 3 has larger QoS weight. On the other hand, if the capacity allocation is optimized to maximize the uplink sum rate, the fronthaul capacity allocated to eRRH 3 is the smallest.

3.6 Summaries, Discussions and Outlooks

In this chapter we have investigated the optimal network design for the uplink transmission of C-RAN and F-RAN. As many existing works have already indicated, the core difficulty for uplink transmission is how to exploit the limited fronthaul capacity resources. The signal compression is one of the key techniques to deal with this problem. Hence, we mainly focus on designing optimal compression

Figure 3.13: The optimal capacity allocation for maximizing the sum rate.

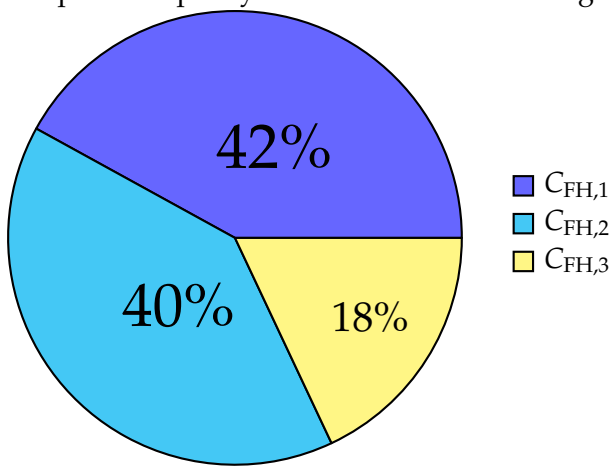
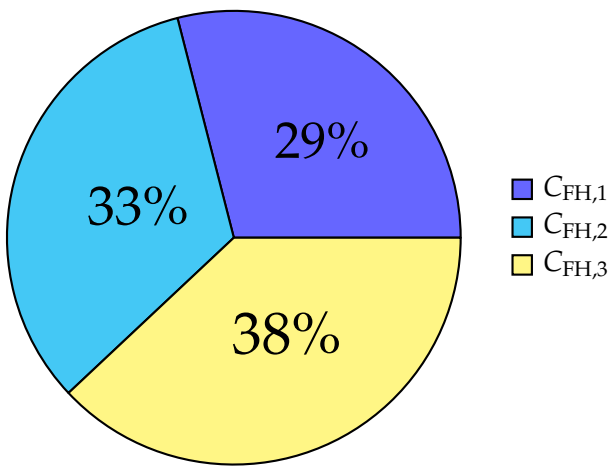


Figure 3.14: The optimal capacity allocation for maximizing the weighted sum rate.



strategies at eRRHs. Furthermore, when the capacity can be shared between different fronthauls, the resource allocation is also worth to be considered for better performance. In order to handle this complicated problem step by step, we firstly assume fixed and known fronthaul capacities, and focus on how to jointly design optimal compression strategies at each eRRH, in order to accommodate to its available resource. Although this problem has been investigated by many works, most of them evaluate it from the perspective of the information theory, by assuming the Gaussian codebook and modeling the compression process by adding artificial additive Gaussian quantization noise. But how shall a practical quantizer that can work for arbitrary codebooks is still left blank. We tackle this problem by extending the conventional IB method and the Bi-Section method to the AIB method and the alternating Bi-Section method. With the proposed AIB method, the trade-off surface for the case of multiple quantizers exploiting the Wyner-Ziv coding can be obtained. And with the alternating Bi-Section method, the specific point on this surface can be efficiently located, at which the corresponding optimal quantizers are acquired, and the fronthaul capacity resources are fully utilized. Then we further consider the scenario where eRRHs share the capacity resource of a common fronthaul. In this case, how to allocate capacity resources to different eRRHs is also a problem, which interacts with the optimal design of the compression strategies for eRRHs. By combining the proposed AIB method and the alternating Bi-Section method with the outer linearization method, we proposed an algorithm to jointly optimize the resource allocation and quantizers. Moreover, we also investigate the network design when different UEs have different QoS targets.

It should be noted that although the proposed algorithms show promising results, the realization of them requires global CSI knowledge at the BBU pool in the cloud. Hence, a large amount of overhead are expected. Moreover, a centralized joint optimization can incur huge computational burden and latency at the BBU pool. Therefore, an outlook for future research direction is a more simplified or distributed mechanism, where even partial CSI is sufficient. Moreover, as we have already stated when introducing this mechanism, although it can be extended to more eRRHs and to multiple antennas conceptually straightforward, the computational complexity, as well as the memory required, will increase exponentially when more and more eRRHs are deployed in the network, or more antennas are mounted on each eRRH. Hence, another promising research direction for the future work might be algorithms based on the concept of the proposed AIB method, but with lower complexity.

Chapter 4

Centralized Joint Design for the Downlink

This chapter contains

[1] After considering the uplink transmission, this chapter focus on the design for the downlink. Nowadays, the downlink of most existing networks requires much higher data rate than the uplink. This is mainly due to the fact that most people acquire much more information via the network through the downlink transmission compared with the information shared by them via the uplink. Hence, a proper design of the downlink contributes more to the overall performance of the network. When we talk about the downlink transmission of F-RAN, as shown in Fig. 4.1, we mean the overall procedure that the scheduled UEs acquire contents from the core network. As we only focus on the F-RAN, we consider only the transmission part starting from the BBU pool, until the UE end. As shown in the figure, the BBU pool sends the requested contents, which have already been processed in the cloud, to eRRHs via fronthauls. Then the eRRHs can execute some further signal processing steps on the received signals, with its fog computing capability. At last the eRRHs transmit the processed signals to UEs, and the scheduled UEs then receive and decode the contents that are intended to themselves. Note that only the scheduled UEs can be active in a specific downlink slot. Similar to the uplink, they have already been notified to be scheduled by the DL-DCI via PDCCH, before the downlink data transmission starts.

[1] Parts of this chapter have been published in [CK16a; CK16f; CSK16; CK17a; CK17b; CK17c; Che+18].

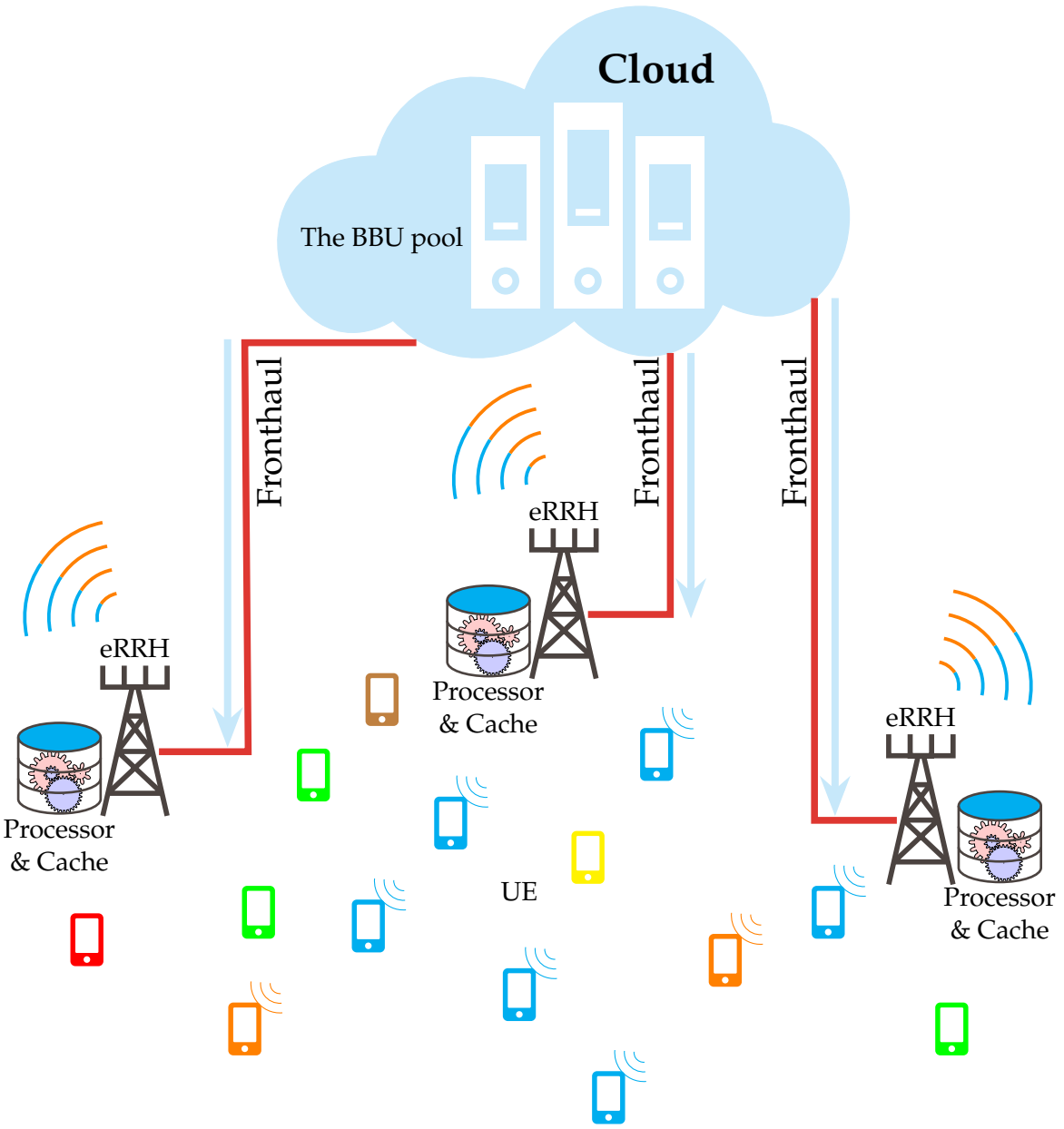


Figure 4.1: The downlink transmission of F-RAN, which consists of eRRHs with both signal processing and storage capabilities. UEs that are receiving signals are scheduled. Two multi-cast groups (depicted in orange and blue) are formed in this specific downlink slot. The content requested by the multi-cast group depicted in blue has already been cached at eRRHs, but the other requested content has to be fetched via the fronthaul.

In this chapter, we are going to investigate the optimal design for the downlink transmission of F-RAN. Similar to the uplink case investigated in the previous chapter, the network design and optimization of the downlink are also done centrally at the BBU pool in the cloud. Similarly, the capacity-limited fronthaul is still a significant bottleneck for the downlink. However, the downlink scenario is more complicated. For example, in the uplink, as each eRRH receives a superposed signal from all scheduled UEs, the design for the compression strategy is the main issue that shall be considered. However, as we are going to show in next sections, there are much more issues required to be considered in the downlink. As the fronthaul capacity is limited, the BBU pool has the freedom to decide, for a specific requested content, to which eRRHs it shall be sent. When more eRRHs receive this content and participate in the transmission of it, more capacity resources are needed, but better performance can be realized: From the viewpoint of the BBU pool, higher aggregate array gain when transmitting this content is realized. As stated in Section 1.3, there are various approaches considered to realize the fog computing capability at eRRHs, one of them is to equip each RRH with a cache module, which is an easy and low-cost way to reduce the requirements on fronthauls. Such a network is called the cache-enabled F-RAN [Tao+16], and it is the main topic we are going to investigate in this chapter. In such a system, at each eRRH, some popularly requested contents can be downloaded and locally cached at the *off-peak* time. When UEs request such contents, they are not necessary to be fetched remotely again and again from the cloud. Hence, the burden on fronthauls can be relieved, and its capacity is not the performance bottleneck for UEs requesting these contents. Moreover, the overall latency might also be greatly reduced. In this case, via the functional split, the eRRHs need to implement certain amounts of signal processing functionalities, which are executed at the BBU pool in the conventional C-RAN. Obviously, this strategy is efficient and can achieve some benefits only when the popularity of different contents varies significantly for most UEs. Some recent predictions and research results from both industry and academia [Cis12; Int12; Sha+13] indicate that, multimedia streaming services will generate a significant portion of the traffic in 5G. For example, some newly released HD Clips or live sport matches might be rather popular in some specific periods of time. When these contents are requested by many UEs simultaneously, a multi-cast scenario is thus formed. Obviously, the larger memory size the cache module has, the more contents it can store, and the transmission burden on fronthauls can thus be relieved more significantly. In particular, the conventional C-RAN can also be regarded as a special version of it, but with the cache memory size of 0. Hence, we can focus only on the problem of the network optimization for the downlink of cache-enabled F-RAN, whose results can be simply applied and extended to the conventional C-RAN, via setting all cache-related items to be 0.

Concerning the issue of adopting the concept of cache for the realization of F-RAN,

several totally different topics are worth to be investigated. For example: What to cache; When to cache; Where to cache and How to cache. These topics can be generalized to the cache placement problem. There have been sufficient works addressing such a problem, in which the efficiency of different caching strategies from the perspective of the information theory, and how to distribute the contents at different eRRHs, so as to fulfill different objective criteria, are investigated. With the pioneering work [MN14], the upper and lower bounds of the capacity for the caching system are characterized. Moreover, the fact that the network capacity can be further increased with the coded multi-casting, is mathematically proved. In their work, two fundamental caching strategies are proposed : Uncoded caching and coded caching. With the uncoded caching strategy, complete files are cached everywhere (in our F-RAN scenario, it denotes all eRRHs). While with the coded caching strategy, distinct fractions (e.g. parity bits) of the same file can be cached at different places (in our F-RAN scenario, it denotes different eRRHs). The fractions can be obtained by using MDS codes (e.g. Fountain code). Considering the problem of content distribution, several schemes are proposed and investigated in [DY16a; Liu+17].

As we mainly focus on the optimal design of the network, we will not go deeply into the cache placement problem in this work. From the perspective of the RAN design, besides the cached contents that are requested, the delivery of the non-cached contents also need to be investigated, as the transmission of them still consumes resources of the network, although the burden on the fronthaul can be greatly released due to the existence of the cache. Hence, under this topic, one important issue for the downlink is the fronthauling strategies for transmission, i.e., how to deliver the requested contents, that are not cached at eRRHs, from the cloud to UEs. Or in another word, how to achieve the optimal downlink performance of the F-RAN, with the help of the cache modules equipped at eRRHs. In order to answer this question, different downlink transmission strategies are proposed and investigated, for meeting different performance criteria for the network. Before we go deeply into this issue, we firstly give a short overview to the state-of-the-art: Regarding the problem of the downlink content delivery, i.e., how to deliver the requested contents from the core network to the UE, via fronthauls, eRRHs and the radio links, there are basically two transmission strategies up to now: The **data-sharing strategy**, or in another word, the **hard transfer mode**, and the **compressed-based strategy**, or in another word, the **soft transfer mode**. The details will be introduced in the next section. For the conventional C-RAN, they are introduced and studied in [PDY15; DY16b]. In these works, only the minimal network energy consumption of these two strategies are compared. However, some important issues for the C-RAN, e.g., the traffic scheduling on fronthauls, and the resultant clustering manner, is not intensively discussed. In [PSS16], the soft and hard transfer modes are compared for the F-RAN. Particularly, how to maximize the minimum-user achievable rate

is studied. For the radio access hop, i.e., the transmission from the RRHs to UEs, the beamforming strategies are well investigated in [DY14; SZL14; DY15; Tao+16; UAS16]. It should be noted that all literature listed above and most existing works assume perfect CSI knowledge available at the BBU pool in the cloud, and based on which the network design is executed. However, in the practical implementation of C-RAN or F-RAN, the CSI are usually estimated and collected at each RRH or eRRH at the network edge, such information are then compressed and delivered to the cloud. Therefore, the distortion of CSI is inevitable, and the introduced CSI error is unknown.

Besides the issues introduced above, a recent report [Cla19] shows that for a typical 5G base station, 300% to 350% more electricity power can be consumed, compared with the energy consumption of a typical LTE base station! Moreover, due to the much denser deployment of the 5G base stations, the total energy consumption of a specific 5G RAN can be unimaginable. A recent reports [New20] states that, China Unicom, one of the biggest 5G network operator in the world, even put many 5G base stations into the sleep mode overnight, in order to save energy. Therefore, the issue of the energy consumption, might be almost the same significant as the issue of the throughput for 5G RAN. According to 3GPP TS 38.211 [3GP18], the number of the slots allocated to the downlink is usually a multiple of that allocated to the uplink. Hence, as stated before, a proper design of downlink can dominate the overall performance of the network to achieve specific criteria, no matter whether we would like to achieve a RAN with maximal throughput at peak times with QoS requirements, or to achieve a *greener* network at off-peak times or scenarios with limited power supply.

Therefore, in this chapter, both high Energy Efficient (EE) oriented network design, targeting at minimizing the energy consumption, while the required QoS can still be guaranteed, and high Spectral Efficient (SE) oriented network design, focusing on higher throughput or balanced QoS, will be investigated. In either network design, we manage to fill some gaps between the existing mechanisms and some critical issues still left blank for the practical implementation. For example, we investigate the issue of the traffic load balancing between different fronthauls, according to their individual available capacities, and the resultant eRRH clustering manner, i.e., which eRRHs shall serve which UEs. Considering the high EE oriented design, not only the power consumption for the transmission is to be taken into account, but also the additional operational power when a RRH or eRRH is actively serving UEs. Therefore, in several circumstances, it might be better to switch off some RRHs or eRRHs so as to save more energy, and the rest ones can still fulfill the UEs' requirements. Thus, we are going to propose an mechanism, which can optimally select which RRH or eRRH can be switched off. Furthermore, the case when only inaccurate CSI knowledge is available will also be addressed. In this scenario, an

algorithm which can robustly design the network is proposed, with which the requested QoS can still be guaranteed for each UE. Additionally, similar to the case of uplink in the previous chapter, the optimal fronthaul resource allocation for the downlink will also be discussed, when it is shared by multiple eRRHs. Here we emphasize again that we will not address the problem of cache placement in this work, since our main focus is the real-time network design and optimization. The cache placement is usually non real-time and done at the off-peak time [Sha+13; Wan+14]. Moreover, we assume the well-known uncoded caching scheme [MN14] is adopted in the network, i.e., all eRRHs cache the same complete contents, which is simple to be executed in practice.

4.1 System Model

4.1.1 Overview

As we stated before, the C-RAN can be regarded as a special case of F-RAN, so it is sufficient to consider the F-RAN model depicted in Fig. 4.1. The BBU pool in the cloud intends to multi-cast[2] different contents to different UE groups via the downlink of F-RAN. The requested data contents, that are not cached at eRRHs, are fronthauled via the above-mentioned hard or soft transfer mode to the network edge. After receiving these contents, the eRRHs will firstly perform several specific signal processing procedures, which will be introduced later in detail, and then send the processed signals further to UEs via the radio access channel.

More specifically, the BBU pool connects to N eRRHs via fronthauls. The capacity of the fronthaul from BBU pool to eRRH $n \in \mathcal{N} = \{1, 2, ..., N\}$ is denoted by $C_{\mathrm{FH},n}$. Each eRRH is equipped with L antennas and a cache module. In each downlink slot, K single-antenna UEs, which are uniformly and independently distributed within the network, are scheduled. The BBU pool knows which content is requested by which scheduled UE in advance [3]. Let M denote the number of distinct contents being requested, UEs requesting the same content (depicted in the same color in Fig. 4.1) form a multi-cast group. In particular, if UE k requests content f^m, $m \in \mathcal{M} = \{1, 2, ..., M\}$, it is classified to multi-cast group $\mathcal{G}^m$, i.e., $k \in \mathcal{G}^m$. We assume that each UE can request at most one content at its scheduled downlink slot. Hence, for any $i, j \in \mathcal{M}$, $\mathcal{G}^i \cap \mathcal{G}^j = \emptyset$, $\forall i \neq j$, and $\sum_{m=1}^{M} |\mathcal{G}^m| \leq K$ must hold. The m-th multi-cast group $\mathcal{G}^m$ is cooperatively served by a *cluster* of eRRHs, denoted by $\mathcal{C}^m$

[2] The uni-cast scenario can also be regarded as as special case of the multi-cast: Each multi-cast UE group consists of only one UE.

[3] The requirements of each UE should have been already sent to the BBU pool in previous uplink slots, the time interval between the uplink and corresponding downlink slot can refer to 3GPP TS 38.211 [3GP18].

with $\mathcal{C}^m \subseteq \mathcal{N}$. Unlike the multi-cast group $\mathcal{G}^m$, which is predetermined and known to the BBU pool based on UEs' requests, the cluster $\mathcal{C}^m$, $\forall m \in \mathcal{M} = \{1, 2, ..., M\}$ is to be dynamically optimized, and the clusters can overlap with each other, in another word, a eRRH can serve several multi-cast groups by delivering different requested contents simultaneously, i.e., for any $i, j \in \mathcal{M}$, $\mathcal{C}^i \cap \mathcal{C}^j$ is not necessary empty.

4.1.2 Content and Cache Model

Firstly we give an introduction to the content model adopted in this work. In a downlink slot, totally M distinct contents are assumed to be requested among M_{total} contents. Each of them is supposed to have a normalized size and is available at the BBU pool in the cloud. However, they might have different probabilities (i.e., popularity) to be requested by the scheduled UEs. Without loss of generality, the requested contents are sorted in the order from the most to the least popular with indices, i.e., the most probable requested content is tagged with index f^1, and the least one is with index f^m. The popularity is modeled by the well-known Zipf distribution [Sha+13], which is widely used in many works: The probability that content f^m is requested can be expressed as

$$\Pr(f^m) = \frac{m^{-\alpha}}{\sum_{j=1}^{M} j^{-\alpha}}, \qquad m = \{1, 2, ..., M_{\text{total}}\}. \tag{4.1}$$

Parameter α is related to the skewness of the distribution, larger α indicates a more biased popularity distribution.

Now we introduce the cache model. Similar to [PSS16; Tao+16], the uncoded caching scheme is adopted, i.e., each content is stored in its original form without coding or multiplexing with other contents. Let integer $S_n \in \mathbb{N}$ denote the cache memory at the n-th eRRH. Each cache module stores the contents according to its popularity until the memory is full. Hence, contents with indices smaller than or equal to S_n will be cached at eRRH n. Let $c_n^{f^m} \in \{1, 0\}$ indicate whether content f^m, which is requested by multi-cast group $\mathcal{G}^m$, is cached at eRRH n or not, i.e.,

$$c_n^{f^m} = \begin{cases} 1 & \text{content } f^m \text{ is cached at eRRH } n, \\ 0 & \text{content } f^m \text{ is not cached at eRRH } n. \end{cases} \tag{4.2}$$

Obviously, for C-RAN, we have $c_n^{f^m} = 0 \; \forall \; m, n$. The cached contents are assumed to be predetermined and known to the BBU pool, as the caching procedure has been completed at the off-peak time. The requested contents that are not cached are firstly transmitted from the BBU pool to eRRHs via fronthauls, then all requested contents are sent to UEs by eRRHs via multi-casting. Compared with the coded

caching scheme, as introduced in Subsection 1.3.1, where different eRRHs cache different fractions of a file, the uncoded caching has lower content diversity but can achieve higher spatial diversity by the cooperative transmission of the same content, which potentially leads to less power consumption. However, the drawback is the higher burden on fronthauls, as the uncached contents has to be fronthauled to multiple eRRHs. Hence, the traffic load handling is a significant issue especially for the uncoded caching. This issue has not been intensively addressed in the previous works concerning F-RAN. After the UEs submit their content requests according to the Zipf distribution (4.1), then at the downlink slots, the cached contents are transmitted directly from eRRHs without consuming the fronthaul resources. Contrarily, all uncached contents being requested must be fetched remotely, from the cloud to each eRRH of the cluster serving the corresponding multi-cast group.

4.1.3 Power Model

In [Ae11], it is shown that for a typical micro base station in LTE, the average transmission power is usually only 6.3 Watt. However, all additional operational power (incl. the power consumed by cooling system, ADC circuits, etc.) can be as high as 56 Watt! Moreover, F-RAN is featured by its relatively large fronthaul capacity, and the fronthaul might also consume considerable power. Hence, the scheme used in [Pen+16; PSS16; Tao+16; UAS16], i.e., activating all eRRHs for higher potential aggregated array gain, so as to decrease the total transmission power of the network, might not necessarily pay off finally: The introduced operational power can be much higher than the saved transmission power. In such scenarios, it is wiser to switch off some eRRHs: Although less cooperative transmission can result in higher transmission power consumption, the saved operational power might compensate it completely. Thus, in order to design greener networks, it is more reasonable to consider the power consumption at the system level, instead of only focusing on how to decrease the transmission power.

Similar to [Ae11; SZL14; TTJ15], the total power consumption of eRRH n is modeled as

$$P_n = \begin{cases} P_{\text{active},n} = \frac{1}{\xi} P_{\text{TX},n} + P_{\text{o}}, & \text{when } P_{\text{TX},n} > 0, \\ P_{\text{sleep}}, & \text{when } P_{\text{TX},n} = 0, \end{cases} \tag{4.3}$$

where $P_{\text{TX},n}$ denotes power consumed by transmission. It is assumed to be limited by the maximal transmission power $P_{n,\max}$. The power amplifier efficiency is denoted by $\xi \in (0,1)$. When eRRH n is activated, i.e., $P_{\text{TX},n} > 0$ holds, its fronthaul and itself are in *active* mode. Let P_{o} denote all additional operational power consumed by an active eRRH. When eRRH n is deactivated, it is in *sleep* mode and does not serve UEs, thus $P_{\text{TX},n} = 0$. P_{sleep} is usually much lower than P_{o}.

4.1.4 Fronthauling Strategies

As stated before, the requested contents that are not cached at eRRHs are to be conveyed through fronthauls. In general, the fronthaul resource allocation can be classified into two categories: *dedicated* and *non-dedicated*. As introduced in Chapter 1, when the fronthaul is constructed, e.g., using the optical-fiber, the capacity of each fronthaul, i.e., the capacity between each eRRH and the BBU pool, is fixed. In this case, the fronthaul transmissions are usually wired communication, e.g., the optical fiber communication. Each eRRH does not need to share the capacity with others. Obviously, the dedicated fronthaul can provide high-capacity communication to each eRRH. However, the price of it is its extreme high cost, especially when micro eRRHs are densely deployed in some urban areas. Moreover, the dedicated fronthaul makes the deployment of eRRHs rather inflexible. As stated in [DC15], for dense and heterogeneous network, such a dedicated fronthaul is not feasible. A counterpart of it can be called as the non-dedicated fronthaul. From the name, we can easily get the point that eRRHs are sharing the total available capacity. As studied in [DC15] and [Gre+15], the wireless fronthaul can be a specific realization of such a sharing manner. Compared with its dedicated brother, the non-dedicated fronthaul can have much lower costs, and the deployment of eRRHs can be rather flexible. Hence, for the dense and heterogeneous network, the non-dedicated fronthaul with wireless communication seems to be the only choice [DC15]. Everything has a price, the price for the non-dedicated fronthaul is that, eRRHs contest with each other for limited capacity resources, thus an efficient capacity resource allocation mechanism between eRRHs is necessary. A straightforward mechanism might be Time-Division (TD) or Frequency-Division (FD) of the resources. Although they are not optimal from information theoretical point of view, the practical implementation of them is rather simple and the cost is low. When the non-dedicated fronthaul is discussed in this chapter, we suppose such an orthogonal implementation. For different design targets, e.g., high EE or high SE, the corresponding optimal resource allocation schemes are worth to be investigated.

Another interesting issue related to the fronthaul transmission is, how to deliver the requested contents that are not cached at eRRHs. As introduced before, there are mainly two fronthauling strategies in general, we call them the hard transfer mode and the soft transfer mode. Briefly, the difference between them mainly lies in how to split the signal processing functionalities between the cloud and the network edge:

- When the **hard** transfer mode is adopted, nearly all signal processing procedures executed on the requested contents are performed at eRRHs on the network edge. The BBU pool only needs to guarantee the reliable transmission of the raw data streams from the cloud to eRRHs;

- When the **soft** transfer mode is adopted, most signal processing procedures executed on the requested contents are performed directly at the BBU pool, including even the modulation step. The resultant signals are then compressed and delivered to the network edge. The eRRHs only decompress the received signals and send them without further processing.

When cache modules are equipped at eRRHs, the soft transfer mode introduced above becomes more complicated, as the requested contents that are locally cached at eRRHs cannot be processed at the BBU pool. In this case, the eRRHs have to undertake the execution of all corresponding processing steps on them.

Before we deep into these two modes, several expressions for easier explanation must be introduced: At the BBU pool, let s^m be a transmitted symbol of content f^m, it has normalized power $\mathbb{E}\{|s^m|^2\} = 1 \; \forall m \in \mathcal{M}$.

4.1.4.1 Hard Transfer Mode

In Fig. 4.2, the abstract model of the downlink transmission is illustrated, when the hard transfer mode is adopted. In this case, the raw data streams of different contents, that are not cached at eRRHs, are firstly encoded (the Gaussian codebook is assumed for simplicity) **separately** and **independently** at the BBU pool, then the encoded raw data streams are sent to different subset of eRRHs (clusters). Besides the encoding step to ensure the reliable delivery of these contents to eRRHs, nothing more is to be implemented at the BBU pool. As stated before, UEs in multicast group $\mathcal{G}^m$ are served by eRRHs in cluster $\mathcal{C}^m$ with $\mathcal{C}^m \subseteq \mathcal{N}$. A specific eRRH might be involved in several clusters, the multi-cast groups $\{\mathcal{G}^m\}_{m=1}^M$ are fixed and known while eRRH clusters $\{\mathcal{C}^m\}_{m=1}^M$ are to be optimized. In more detail, when eRRH $n \in \mathcal{N}$ is involved in several clusters, i.e., it is responsible for transmitting the several requested contents, **Function Block ENC n** at the BBU pool in Fig. 4.2 represents the parallel and independent encoding of these requested but uncached contents, which eRRH n should be responsible for. Then these encoded data streams are transmitted via the fronthaul of capacity $C_{\mathrm{FH},n}$ to eRRH n. At eRRH n, **Function Block DEC n** represents the parallel and independent decoding of these data streams. Here we assume an error-free fronthaul transmission for simplicity. Afterwards, together with the locally cached requested contents, **Function Block BF n** undertakes the corresponding beamforming of all contents that are transmitted by eRRH n. In the end, the beamformed data streams are multiplexed with each other and modulated into the appropriate signal, via **Function Block MUX n** and **Function Block MOD n**, respectively.

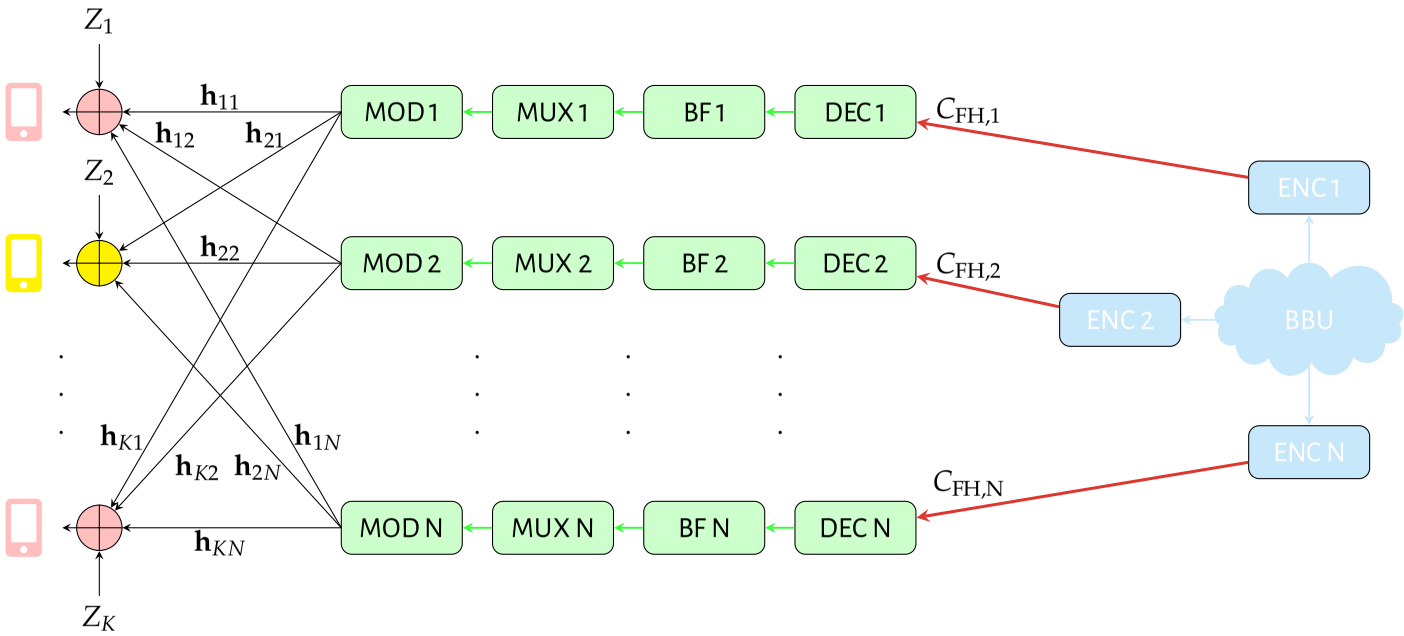


Figure 4.2: The abstract model of the downlink multi-cast F-RAN adopting the **hard** transfer mode and dedicated fronthaul. **Function Block ENC n**: Encoding of the uncached raw data streams for eRRH n at the BBU pool; **Function Block DEC n**: Decoding of the uncached raw data steams at eRRH n; **Function Block BF n**: Beamforming of all data streams with the corresponding beamformers at eRRH n; **Function Block MUX n**: Multiplexing of all beamformed data streams at eRRH n; **Function Block MOD n**: Modulation of the multiplexed data at eRRH n. UEs depicted in the same color indicate that they request the same content.

Note that eRRHs in the same multi-cast cluster form a distributed MIMO system and allow Cooperative Multi-Point (CoMP) transmission. Let $\mathbf{v}^m = [\{\mathbf{v}_1^m\}^H, \{\mathbf{v}_2^m\}^H, ..., \{\mathbf{v}_N^m\}^H]^H \in \mathbb{C}^{NL\times 1}$ denote the aggregated beamformer constructed among all eRRHs for content f^m, where $\mathbf{v}_n^m = [v_{n,1}^m, v_{n,2}^m, ..., v_{n,L}^m]^T \in \mathbb{C}^{L\times 1}$ indicates the beamformer constructed at eRRH n. If eRRH n is not involved in cluster $\mathcal{C}^m$, we have $\mathbf{v}_n^m = \mathbf{0}$, or equivalently, the ℓ_0-norm of its power is 0, i.e., $\left|||\mathbf{v}_n^m||_2^2\right|_0 = 0$. Otherwise it is a non-zero vector, and the ℓ_0-norm of its power is 1. When content f^m is delivered to UEs with rate R^m, then for the **dedicated** fronthaul, the capacity requirement of the fronthaul connected to eRRH n, i.e., $C_{\text{req},n}^{\text{hard}}$, must satisfy

$$C_{\text{req},n}^{\text{hard}} = \sum_{m=1}^{M} (1 - c_n^{f^m}) \left|||\mathbf{v}_n^m||_2^2\right|_0 R^m \leq C_{\text{FH},n} \qquad \textit{dedicated}. \tag{4.4}$$

Specifically, when eRRH n is involved in cluster $\mathcal{C}^m$, we must have $\left|||\mathbf{v}_n^m||_2^2\right|_0 = 1$. If content f^m is not cached, then we have $c_n^{f^m} = 0$. Only in this case, i.e., eRRH n shall transmit f^m and this content is not cached at eRRH n, at least rate R^m for content f^m has to be supported by the fronthaul between the BBU pool and eRRH n. By summing up all contents in $\mathcal{M}$, we can achieve the inequality above.

Similarly, for the **non-dedicated** fronthaul, the following constraint needs to be satisfied:

$$C_{\text{req},n}^{\text{hard}} = \sum_{n=1}^{N} \sum_{m=1}^{M} (1 - c_n^{f^m}) \left|||\mathbf{v}_n^m||_2^2\right|_0 R^m \leq C_{\text{FH}} \qquad \textit{non-dedicated}. \tag{4.5}$$

Here, C_{FH} denotes the total capacity of the fronthaul to be shared.

Via optimizing the beamformers, as shown later, the cluster for the hard mode can also be optimized, and traffic on fronthauls are scheduled according to their individual capacities. Moreover, we can become aware of whether switching off some eRRHs is beneficial.

As stated before, when a specific content can be delivered to more eRRHs, i.e., the corresponding multi-cast cluster becomes larger, the transmission power for this content can be reduced, or the QoS of this content can be improved, due to higher spatial diversity gain. But it also consumes more fronthaul resources, and might lead to an otherwise sleeping eRRH to be active again, which further results in more operation power consumption. Hence, there is an interaction between the transmission power, the operation power, the QoS target, and the available fronthaul capacity.

4.1.4.2 Soft Transfer Mode

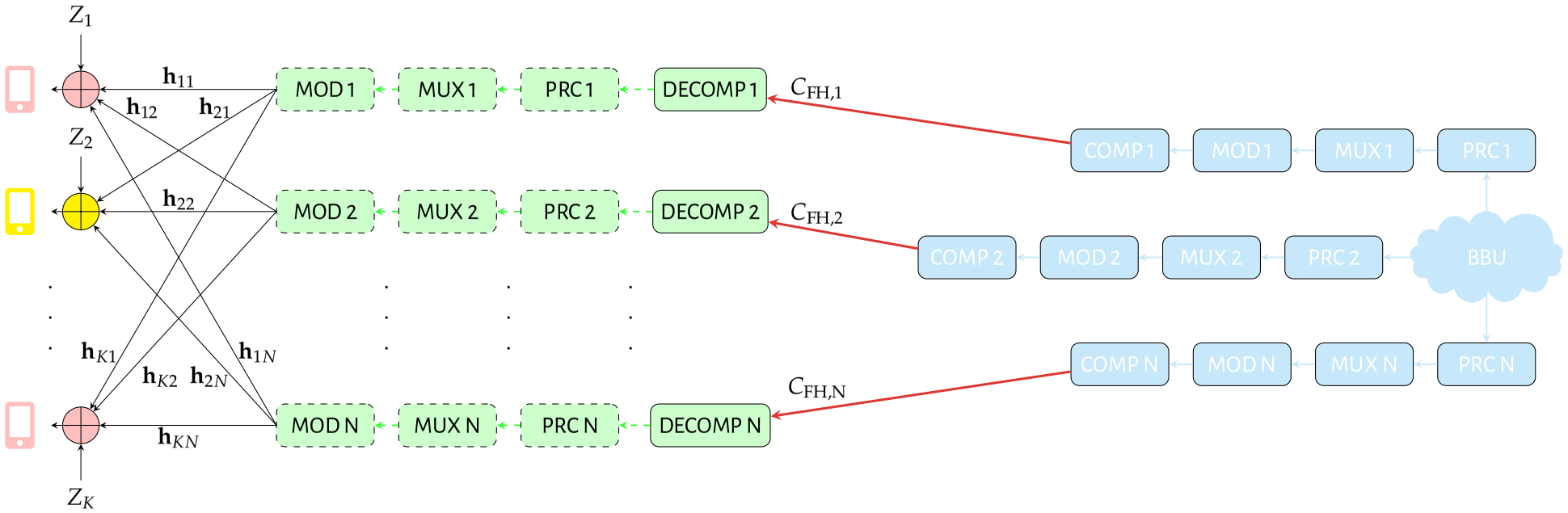


Figure 4.3: The abstract model of the downlink multi-cast F-RAN adopting the **soft** transfer mode and dedicated fronthaul. **Function Block PRC n**: Precoding of the requested but uncached raw data streams for eRRH n at the BBU pool; **Function Block MUX n**: Multiplexing of the precoded uncached data streams for eRRH n at the BBU pool; **Function Block MOD n**: Modulation of the multiplexed data for eRRH n at the BBU pool; **Function Block COMP n**: Compression of the modulated signal (incl. encoding of the compression indices) for eRRH n at the BBU pool; **Function Block DECOMP n**: Decompression and reconstruction of the modulated signal (incl. decoding of the compression indices) at eRRH n; **Dashed Function Block PRC n**: Precoding of the requested and locally cached contents at eRRH n. **Dashed Function Block MUX n**: Multiplexing of the precoded cached contents at eRRH n; **Dashed Function Block MOD n**: Modulation of the multiplexed data resultant from the locally cached contents at eRRH n. All dashed function blocks affect only on locally cached contents, and are not valid for the uncached contents that are soft-fronthauled from the cloud. UEs depicted in the same color indicate that they request the same content.

As introduced in the previous subsection, when the hard transfer mode is adopted, most signal processing procedures are executed at the network edge. However, when the soft transfer mode is adopted, it is the other way around: If some contents have to be fetched from the cloud, i.e., they are requested but not cached at eRRHs, the precoding, multiplexing, as well as the modulation are to be applied on them directly at the BBU pool. The modulated signal have to be then compressed, in order to satisfy individual fronthaul capacity constraints. The compressed signals are sent to the corresponding eRRHs. At eRRHs, besides the reconstruction of the received compressed signals, no further processing shall be implemented on the uncached contents, the reconstructed signals are directly forwarded to UEs. However, for the requested contents that are locally cached at eRRHs, no compression and soft fronthauling are required: They are precoded, multiplexed, and modulated locally before being sent to UEs. An abstract model of the soft transfer mode is illustrated in Fig. 4.3. In more detail, **Function Block PRC n** at the BBU pool represents the parallel and independent precoding of all requested but uncached contents, that shall be transmitted by eRRH $n \in \mathcal{N}$. **Function Block MUX n** represents the multiplexing of the previously precoded data streams for eRRH n at the BBU pool. The resultant data stream is then modulated with **Function Block MOD n**. In the end, **Function block COMP n** compresses the modulated signal for eRRH n, with which the fronthaul of capacity $C_{\mathrm{FH},n}$ can support the transmission of it. Then at eRRH n, **Function Block DECOMP n** represents the decompression and reconstruction of the fronthauled signal. We must emphasize that **Function block COMP n** and **Function Block DECOMP n** also incorporate the encoding and decoding procedure of the compression indices, respectively. Together with **Dashed Function Block PRC n**, **MUX n**, and **MOD n**, which should be applied only on the locally cached contents, eRRH n then transmit all requested contents to UEs.

Let $\mathbf{w}^m = [\{\mathbf{w}_1^m\}^H, \{\mathbf{w}_2^m\}^H, ..., \{\mathbf{w}_N^m\}^H]^H \in \mathbb{C}^{NL\times 1}$ be the aggregated precoders for $\mathcal{G}^m$, where $\mathbf{w}_n^m = [w_{n,1}^m, w_{n,2}^m, ..., w_{n,L}^m]^T \in \mathbb{C}^{L\times 1}$ indicates the precoder intends for eRRH n. After the multiplexing step, the superposed signal $\widetilde{\mathbf{x}}_n$ constructed at the BBU pool for eRRH n is

$$\widetilde{\mathbf{x}}_n^{\mathrm{soft}} = \sum_{m=1}^{M} (1 - c_n^{f^m}) \mathbf{w}_n^m s^m. \tag{4.6}$$

For the modulation step, we consider the ideal Gaussian alphabet with infinite cardinality for simplicity, as the modulation scheme is not the topic investigated in this work. Therefore, the modulation step is supposed not to introduce further distortions to $\widetilde{\mathbf{x}}_n^{\mathrm{soft}}$.

Due to the fronthaul capacity constraints, $\widetilde{\mathbf{x}}_n^{\mathrm{soft}}$ must be compressed before transmission: We assume independent compression procedures for each antenna in this work, i.e., no Wyner-Ziv coding is performed, although this is not optimal from

the information-theoretical perspective, as the correlation between antennas is not exploited, it is a practical solution and causes much less delay and complexity. A joint compression strategy is also studied, details can be found in [Par+13a]. Let $\mathbf{e}_n^{\text{comp}} = [e_{n,1}^{\text{comp}}, e_{n,2}^{\text{comp}}, ..., e_{n,L}^{\text{comp}}]^T \in \mathbb{C}^{L\times 1}$ denote the artificial quantization noise vector for eRRH n. Specifically, $\mathbf{e}_n^{\text{comp}} \sim \mathbb{CN}\left(\mathbf{0}, \text{Diag}\left([q_{n,1}^2, q_{n,2}^2, ..., q_{n,L}^2]\right)\right)$. Namely, $e_{n,l}^{\text{comp}}$ constructed for the l-th antenna of eRRH n is Gaussian distributed with 0 mean and variance $q_{n,l}^2$. The signal to be delivered to eRRH n becomes

$$\mathbf{x}_n^{\text{soft}} = \widetilde{\mathbf{x}}_n^{\text{soft}} + \mathbf{e}_n^{\text{comp}} = \sum_{m=1}^{M} (1 - c_n^{f^m}) \mathbf{w}_n^m s^m + \mathbf{e}_n^{\text{comp}}, \tag{4.7}$$

with the l-th element for antenna l of eRRH n being expressed as

$$x_{n,l}^{\text{soft}} = \widetilde{x}_{n,l}^{\text{soft}} + e_{n,l}^{\text{comp}} = \sum_{m=1}^{M} (1 - c_n^{f^m}) w_{n,l}^m s^m + e_{n,l}^{\text{comp}}. \tag{4.8}$$

By exploiting chain rule and the independence assumption, for the **dedicated** fronthaul, the fronthaul resource consumption of eRRH n, $C_{\text{req},n}^{\text{soft}}$ should satisfy

$$\begin{aligned} C_{\text{req},n}^{\text{soft}} &= I(\mathbf{x}_n^{\text{soft}}; \widetilde{\mathbf{x}}_n^{\text{soft}}) = \sum_{l=1}^{L} I(x_{n,l}^{\text{soft}}; \widetilde{x}_{n,l}^{\text{soft}}) \\ &= \sum_{l=1}^{L} \log_2 \left(1 + \frac{\sum_{m=1}^{M} (1 - c_n^{f^m}) |w_{n,l}^m|^2}{q_{n,l}^2}\right) \leq C_{\text{FH},n} \qquad \textit{dedicated}. \end{aligned} \tag{4.9}$$

It means that as long as (4.9) is satisfied, eRRH n is able to theoretically reconstruct $\mathbf{x}_n^{\text{soft}}$, and further forward it to UEs. And for the **non-dedicated** fronthaul, we can arrive at a similar inequality as follows

$$\begin{aligned} C_{\text{req}}^{\text{soft}} &= \sum_{n=1}^{N} I(\mathbf{x}_n^{\text{soft}}; \widetilde{\mathbf{x}}_n^{\text{soft}}) = \sum_{n=1}^{N} \sum_{l=1}^{L} I(x_{n,l}^{\text{soft}}; \widetilde{x}_{n,l}^{\text{soft}}) \\ &= \sum_{n=1}^{N} \sum_{l=1}^{L} \log_2 \left(1 + \frac{\sum_{m=1}^{M} (1 - c_n^{f^m}) |w_{n,l}^m|^2}{q_{n,l}^2}\right) \leq C_{\text{FH}} \qquad \textit{non-dedicated}. \end{aligned} \tag{4.10}$$

Obviously, when coarser quantization is to be expected, the quantization noise shall be set larger, i.e., the value of $q_{n,l}$ is increased, less fronthaul resources are to be consumed. However, more distortions are introduced to the final signals delivered to UEs.

4.1.5 Signal Processing at eRRH

As stated in the previous subsection, when the hard transfer mode is adopted, each eRRH can obtain the raw data of the uncached contents that are fronthauled from

the cloud, then they are encoded and beamformed before further sending to UEs. The cached contents are encoded and beamformed directly at the corresponding eRRHs. The beamformers are delivered to each eRRH via pilot signals transmitted by fronthauls, thus certain dedicated fronthaul capacity consumption must be taken in to account when the available fronthaul capacities $C_{\mathrm{FH},n}/C_{\mathrm{FH}}$ in (4.4) and (4.5) is calculated, i.e., the dedicated fronthaul capacity used for pilots has to be deducted from the total available capacity, in order to obtain the values of $C_{\mathrm{FH},n}/C_{\mathrm{FH}}$. The transmitted signal from eRRH n for the multi-cast group $\mathcal{G}^m$ that requests content f^m is

$$\mathbf{x}_n^{m,\mathrm{hard}} = \mathbf{v}_n^m s^m. \tag{4.11}$$

By summing up the signals for all multi-cast groups, the transmitted signal constructed at eRRH n can be written as

$$\mathbf{x}_n^{\mathrm{hard}} = \sum_{m=1}^{M} \mathbf{x}_n^{m,\mathrm{hard}} = \sum_{m=1}^{M} \mathbf{v}_n^m s^m. \tag{4.12}$$

Before the signal is sent from eRRH n to UEs, the following transmission power constraint has to be satisfied (Note that symbol $s^m\ \forall m \in \mathcal{M}$ has normalized power)

$$P_{\mathrm{TX},n}^{\mathrm{hard}} = \sum_{m=1}^{M} ||\mathbf{v}_n^m||_2^2 \leq P_{\mathrm{TX},n}^{\mathbf{max}}. \tag{4.13}$$

When the soft transfer mode is adopted, for all $n \in \mathcal{N}$, eRRH n decompresses and reconstructs the received signal $\mathbf{x}_n^{\mathrm{soft}}$ according to (4.7). As stated before, it only consists of the information that are not cached at eRRHs. For content f^m that is cached at eRRH n, i.e., $c_n^{f^m} = 1$, signal $\mathbf{w}_n^m s^m$ shall be constructed locally at eRRH n. Similarly to the hard transfer mode, the precoders for the cached contents are also delivered to each eRRH via pilot signals transmitted by fronthauls. Hence, certain dedicated fronthaul capacity shall be reserved similarly, when the value of $C_{\mathrm{FH},n}/C_{\mathrm{FH}}$ in (4.9) and (4.10) are computed. After the signals are reconstructed (for the non-cached contents) or constructed (for the cached contents), the eRRH sends them further to UEs. Although the recovered signals will not be further processed, the transmission power constraints of each eRRH still need to be respected when designing the precoders:

$$\begin{aligned} P_{\mathrm{TX},n}^{\mathrm{soft}} &= \sum_{m=1}^{M} (1 - c_n^{f^m}) ||\mathbf{w}_n^m||_2^2 + ||\mathbf{q}_n||_2^2 + \sum_{m=1}^{M} c_n^{f^m} ||\mathbf{w}_n^m||_2^2 \\ &= \sum_{m=1}^{M} ||\mathbf{w}_n^m||_2^2 + ||\mathbf{q}_n||_2^2 \leq P_{\mathrm{TX},n}^{\mathbf{max}}, \end{aligned} \tag{4.14}$$

with $\mathbf{q}_n = [q_{n,1}, q_{n,2}, ..., q_{n,L}]^T$.

Remark: Note that when the soft transfer mode is adopted, only the requested contents that are not cached are distorted, and such a distortion introduces extra transmission power consumption, as shown in the second term of (4.14). Although some power is wasted compared with the hard transfer mode, it can save certain fronthaul capacity resources: Only the precoders for the cached contents need to be transmitted via pilots using dedicated capacities. However, for the hard transfer mode, all beamformers have to be transmitted via pilots.

4.1.6 Radio Access Channel

Usually, the CSI-RS (Channel State Information - Reference Signal) is sent via the downlink to each UE for estimating the channel quality. Then the UEs will feedback the CSI to eRRHs and the BBU pool via the PUCCH (Physical Uplink Control Channel) in uplink slots [3GP18]. Hence, we can suppose the channel information is always available to the BBU pool. Let $\mathbf{h}_n^k = [h_{n,1}^k, h_{n,2}^k, ..., h_{n,L}^k]^T \in \mathbb{C}^{L\times 1}$ be the actual downlink channel vector between eRRH n and UE k. Thus, the aggregated actual downlink channel vector from all eRRHs to UE k can be written as $\mathbf{h}_k = [\mathbf{h}_1^{k\,H}, \mathbf{h}_2^{k\,H}, ..., \mathbf{h}_N^{k\,H}]^H \in \mathbb{C}^{NL\times 1}$. When perfect global CSI is assumed to be known, we say $\{\mathbf{h}_k\}_{k=1}^K$ is available at the BBU pool. Hence, when the hard transfer mode is adopted and Gaussian alphabet with infinite cardinality is assumed for the modulation step, the SINR at UE k can be expressed as

$$\mathrm{SINR}_k^{\mathrm{hard}} = \frac{|\mathbf{h}_k^H \mathbf{v}^m|^2}{\sum_{i\neq m}^M |\mathbf{h}_k^H \mathbf{v}^i|^2 + \sigma_k^2}, \quad k \in \mathcal{G}^m, \tag{4.15}$$

where σ_k^2 denotes variance of the i.i.d additive complex Gaussian noise with zero mean at UE k. From (4.15), the desired signal of each UE is interfered by other uninteresting signals as well as the additive white Gaussian noise.

Similarly, when the soft mode is adopted, the SINR at UE k is

$$\mathrm{SINR}_k^{\mathrm{soft}} = \frac{|\mathbf{h}_k^H \mathbf{w}^m|^2}{|\mathbf{h}_k^H \mathbf{q}|^2 + \sum_{i\neq m}^M |\mathbf{h}_k^H \mathbf{w}^i|^2 + \sigma_k^2}, \quad k \in \mathcal{G}^m, \tag{4.16}$$

where $\mathbf{q} = [\mathbf{q}_1^T, \mathbf{q}_2^T, ..., \mathbf{q}_N^T]^T$ denotes the aggregated vector of the quantization noise across all eRRHs. Obviously, the desired signal of each UE is also interfered by the quantization noise.

4.1.7 Inaccurate CSI

When inaccurate CSI is considered, we adopt a widely used additive error model [Pon+11; GCW12; NN14; SZL15] to describe the inaccurate CSI as follows:

$$\widetilde{\mathbf{h}}_k = \mathbf{h}_k + \mathbf{e}_k^{\text{CSI}}, \tag{4.17}$$

$$\text{with Pr}\left\{||\mathbf{e}_k^{\text{CSI}}||_2^2 \leq \epsilon_k^2\right\} \geq 1 - \delta_k, \quad \forall k \in \{1, 2, ..., K\}. \tag{4.18}$$

Vector $\mathbf{h}_k \in \mathbb{C}^{NL\times 1}$ represents the inaccurate aggregated channel vector for UE k. We assume only such inaccurate information is available at the BBU pool, and based on which the network is optimized. Vector $\mathbf{e}_k^{\text{CSI}} \in \mathbb{C}^{NL\times 1}$ denotes the aggregated CSI error vector of UE k. It is assumed to be bounded in the spherical region with radius ϵ_k with the probability of at least $1 - \delta_k$. δ_k is said to be the *outage probability*. Similar to many existing works [Pon+11; GCW12; NN14; SZL15], we take such a sphere model, instead of the well-known Gaussian model to describe the error pattern. This is mainly due to the fact that such a model is more general, the Gaussian model can be regarded as a special case of it, as long as $\delta > 0$. Obviously, all algorithms proposed later for this sphere model are valid for the Gaussian model. For UE k $\forall k \in \{1, 2, ..., K\}$, the BBU pool knows only the value of ϵ_k, instead of the exact aggregated error vector $\mathbf{e}_k^{\text{CSI}}$, hence, the exact channel knowledge $\widetilde{\mathbf{h}}_k$ is not available.

Note that the CSI error also introduces interference to the desired signal. By substituting (4.17) into (4.15) and (4.16), and treating all interference as noise, including the additional one resulting from the inaccuracy of the CSI, the actual achievable *effective* SINR for UE $k \in \mathcal{G}^m$ for both fronthauling strategies can be expressed as

$$e\text{SINR}_k^{\text{hard}}(\mathbf{e}_k^{\text{CSI}}) = \frac{\left|\mathbf{h}_k^H \mathbf{v}^m\right|^2}{\left|\mathbf{e}_k^{\text{CSI}^H} \mathbf{v}^m\right|^2 + \sum_{i\neq m}^{M} \left|\left(\mathbf{h}_k^H + \mathbf{e}_k^{\text{CSI}^H}\right) \mathbf{v}^i\right|^2 + \sigma_k^2}, \tag{4.19}$$

$$e\text{SINR}_k^{\text{soft}}(\mathbf{e}_k^{\text{CSI}}) = \frac{\left|\mathbf{h}_k^H \mathbf{w}^m\right|^2}{\left|\left(\mathbf{h}_k^H + \mathbf{e}_k^{\text{CSI}^H}\right) \mathbf{q}\right|^2 + \left|\mathbf{e}_k^{\text{CSI}^H} \mathbf{w}^m\right|^2 + \sum_{i\neq m}^{M} \left|\left(\mathbf{h}_k^H + \mathbf{e}_k^{\text{CSI}^H}\right) \mathbf{w}^i\right|^2 + \sigma_k^2}. \tag{4.20}$$

The items in the denominator of (4.20) denote the interference resulting from the signal compression, inaccuracy of the CSI, and the contents intended to all other multi-cast groups, as well as the noise. For the hard transfer mode, there is no quantization noise resulting from the compression, as shown in (4.19). We see that the achievable effective SINRs of both transfer modes are functions of the aggregated error vectors $\{\mathbf{e}_k^{\text{CSI}}\}_{k=1}^K$. Here we emphasize again that the BBU pool does not know them. Hence, the exact value of the achievable effective SINRs can not be derived, due to their dependence on random $\{\mathbf{e}_k^{\text{CSI}}\}_{k=1}^K$.

4.1.8 Summary

Up to now, we have introduced the models and the signal processing procedures adopted in the cache-enabled F-RAN. The requirements of the whole systems are derived, for different transfer modes and different fronthaul resource sharing policies. Then in the next step, we are going to propose feasible and low complexity algorithms to optimize the network for different targets, e.g., whether achieving the maximal throughput of the network the top priority, or a greener networks is preferred. In summary, in order to achieve the performance targets of the network, the results of the proposed algorithms must tell each eRRH:

1. *Which UEs shall be served? Or in another word, which contents shall each eRRH transmit?*
2. *If eRRHs have to share the fronthaul resources with others, how much capacity can be assigned to each one?*
3. *For the hard transfer mode, how shall each content be transmitted? Specifically, how shall each eRRH beamform each content?*
4. *For the soft transfer mode, after the uncached precoded contents are decompressed and forwarded , how shall the cached contents be transmitted? Specifically, how shall each eRRH precode the cached contents locally?*
5. *How shall each eRRH allocate its limited power for different multi-cast groups that it serves? Or in another word, how much power shall be allocated to each content?*
5. *Is it possible to deactivate some eRRHs to save power?*

We are going to answer all questions from the next subsection. By comparing (4.4) and (4.5) for the hard transfer mode, (4.9) and (4.10) for the soft transfer mode, it can be seen that the inequality constraints for the dedicated and the non-dedicated fronthaul resources have similar forms. As will be shown in the next subsections, similar techniques can be applied to deal with these two scenarios. Hence, in order to avoid unnecessary repetitions, for each design target, we only select one specific scenario for intensive investigation. For example, when high EE oriented cache-enabled F-RAN is the target, only algorithms and numerical results for the scenario of dedicated fronthaul will be introduced in detail, as the algorithms for the non-dedicated case can adopt the same techniques and, of course, with some straightforward modifications. Naturally, we will elaborate on how such modifications shall happen for other scenarios, after a detailed derivation of the algorithm for a specific scenario is given.

4.2 Joint Optimization for Different Criteria

In this section, the central optimization method for the downlink is to be investigated. We will start with the case when perfect global CSI is available at the BBU pool, then we will proceed to methods dealing with the case when only inaccurate CSI exists. For both cases, we focus on both high Energy Efficiency (EE) oriented design and high Spectral Efficiency (SE) oriented design. When high EE is the target, minimizing the energy consumption of the network is the main objective, of course under the condition that the QoS of each UE can be guaranteed, and the constraints of the network must be satisfied. More specifically, we consider both cases of minimizing only the transmission power, and its extension where the operational power of an active eRRH is taken into account. In the latter case, as we are going to see next, switching off some eRRHs is able to compensate the increasing of the transmission power resulting from less spatial diversity. When high SE is considered, there are two different variations: The first one is to maximize the network multicast throughput, such that the capability for the downlink multi-casting rate of this network can be squeezed to the limit. However, it can happen that some UEs with poor channel conditions never get scheduled. This is mainly due to the fact that, increasing the achievable rate for these *bad* UEs will consume much more resources of the network, comparing to the ones with good channel qualities. Hence, in order to fully utilize the available network resources for the maximization of the overall network throughput, these *good* UEs are prioritized, which definitely leads to unfairness between UEs. The second variation is to achieve the (weighted) Max-Min Fairness between all UEs. In this case, the lowest QoS of each requested content is maximized, in order to achieve the (weighted) fairness between them.

Although the criteria of network design are far different from each other, in each scenario, the problem of the traffic load balancing, cluster formulation and compression etc. will be intensively discussed. Furthermore, we will then investigate how to guarantee the network performance, in terms of the different design metrics listed above, in the presence of only inaccurate CSI.

4.2.1 High EE oriented Design — TX Power Minimization

In this subsection, in order to make the descriptions and the derivations of the algorithms easy to follow, we start with the simplest case: Only minimizing the transmission power (TX power), without considering the possibility to switch off eRRHs to save the additional operational power. Afterwards we extend the scenario by taking the additional operational power into account, and propose a mechanism for minimizing the total power to make the network greener. For both scenarios,

two different fronthauling strategies, i.e., the hard and the soft transfer modes, are investigated separately, as the signal processing procedures of them, as well as the techniques utilized to deal with the optimization of them, are quite different.

4.2.1.1 Design for the Hard Transfer Mode

When the hard transfer mode is adopted by the network based on (4.4), (4.5), (4.13), (4.15) and the previous introductions, the problem can be formulated as follows:

$$\mathcal{P}_{\text{Hard original}}: \quad \min_{\{\mathbf{v}^m\}_{m=1}^M} \quad \sum_{m=1}^{M} ||\mathbf{v}^m||_2^2, \tag{4.21}$$

$$\text{s.t.} \quad \text{SINR}_k^{\text{hard}} \geq \Gamma^m, \quad \forall k \in \mathcal{G}^m, \quad \forall \mathcal{G}^m, \tag{4.22}$$

$$\sum_{m=1}^{M} (1 - c_n^{f^m}) \left| ||\mathbf{v}_n^m||_2^2 \right|_0 R^m \leq C_{\text{FH},n} \quad \forall n \in \mathcal{N} \qquad \textit{dedicated}, \tag{4.23}$$

$$\sum_{n=1}^{N} \sum_{m=1}^{M} (1 - c_n^{f^m}) \left| ||\mathbf{v}_n^m||_2^2 \right|_0 R^m \leq C_{\text{FH}} \qquad \textit{non-dedicated}, \tag{4.24}$$

$$\sum_{m=1}^{M} ||\mathbf{v}_n^m||_2^2 \leq P_{\text{TX},n}^{\max}, \quad \forall n \in \mathcal{N}. \tag{4.25}$$

Eq. (4.21) describes the transmission power consumption of the network, which is the sum of the transmission power consumed for each multi-cast group among all eRRHs. Constraint (4.22) guarantees the QoS of each UE in each multi-cast group, where Γ^m denotes the target SINR of the content requested by $\mathcal{G}^m$ and $\text{SINR}_k^{\text{hard}}$ is defined in (4.15). When each eRRH is assigned with dedicated fronthaul resource, constraint (4.23) guarantees that the traffic on each fronthaul does not exceed its capacity: As stated in the previous section, we use the ℓ_0-norm to denote whether the beamforming vector $\mathbf{v}_n^m$ is a zero vector or not, i.e., if eRRH n is involved in cluster $\mathcal{C}^m$ serving multi-cast group $\mathcal{G}^m$, it is a non-zero vector and thus $\left| ||\mathbf{v}_n^m||_2^2 \right|_0 = 1$ holds, otherwise the ℓ_0-norm is zero. We see that UEs in multi-cast group $\mathcal{G}^m$ consumes the capacity resource of the fronthaul to eRRH n only if the requested content is not cached, i.e., $c_n^{f^m} = 0$, and this eRRH indeed contributes to multi-cast group $\mathcal{G}^m$, i.e., $\left| ||\mathbf{v}_n^m||_2^2 \right|_0 = 1$. In this case, the fronthaul capacity resource consumption for this uncached content can be written as

$$R^m = \log_2 (1 + \Gamma^m) \tag{4.26}$$

at a minimum, when the Gaussian codebook is used. For all the computations from now on, we assume the Gaussian codebook for simplicity unless otherwise stated. By summing up all multi-cast groups, we obtain the total fronthaul resource consumption of eRRH n in (4.23), which should be smaller than its capacity. Similarly, the capacity constraint for the non-dedicated case is expressed in (4.24). Constraint

(4.25) ensures that at each eRRH, the transmission power does not exceed its maximal allowable power.

The descriptions above indicate that the clustering and the beamformers interact with the requested contents, the cached contents, the fronthaul link capacities, the maximal allowable power, and the radio channel conditions between all eRRHs and scheduled UEs. For different scheduling intervals, i.e., different downlink slots, the parameters above (except for the fronthaul capacities and maximal allowable power of each eRRH) change independently and dynamically, thus an efficient optimization scheme is required. Although we do not explicitly optimize the clustering scheme, i.e., which subset of eRRHs shall serve which multi-cast group, it is implicitly optimized and determined by the resulting value $\left|||\mathbf{v}_n^m||_2^2\right|_0$ of the problem.

Then the question becomes how to solve the problem above. As stated in the last part of the previous subsection, a specific fronthaul sharing strategy is to be selected for illustrating the solution of the problem raised above. Here we select the **dedicated** case, i.e., solving the problem consisting of (4.21), (4.22), (4.23) and (4.25). After completing the introduction of the solution, a short description will be given for amending it to the non-dedicated case.

Note that the objective function (4.21) and the LHS of constraints (4.22) and (4.23) are non-convex functions. Moreover, the ℓ_0-norm in (4.23) makes the corresponding function be step-like and similar to a Mixed Integer Non-Linear Programming (MINLP) problem [MFR20]. Hence, this problem is in general non-convex and NP-hard. Then the first step is to develop methods to convexify the original problem.

At first we adopt the Semi-Definite Relaxation (SDR) technique introduced in Subsection 2.3.4 to convexify (4.22). Let $\mathbf{V}^m = \mathbf{v}^m(\mathbf{v}^m)^H$ and $\mathbf{H}_k = \mathbf{h}_k\mathbf{h}_k^H, \forall m, k$, where both $\mathbf{V}^m, \mathbf{H}_k \in \mathbb{C}^{NL\times NL}$ are positive semidefinite matrices. We further define a selection matrix at eRRH n as $\mathbf{J}_n = \text{Diag}\left(\left[\mathbf{0}^H_{(n-1)L\times 1}, \mathbf{1}^H_{L\times 1}, \mathbf{0}^H_{(N-n)L\times 1}\right]\right)$. Therefore, the following expressions can be derived: $||\mathbf{v}^m||_2^2 = \text{tr}(\mathbf{V}^m)$, $||\mathbf{v}_n^m||_2^2 = \text{tr}(\mathbf{V}^m\mathbf{J}_n)$, and $|\mathbf{h}_k^H\mathbf{v}^m|^2 = \text{tr}(\mathbf{V}^m\mathbf{H}_k)$. Then together with (4.15) and (4.26), the original problem (4.21) - (4.25) can be equivalently reformulated as follows

$$\mathcal{P}_{\text{Hard}}: \min_{\{\mathbf{V}^m\}_{m=1}^M} \quad \sum_{m=1}^{M} \text{tr}(\mathbf{V}^m), \tag{4.27}$$

$$\text{s.t.} \quad \Gamma^m \left(\sigma_k^2 + \sum_{i \neq m}^{M} \text{tr}(\mathbf{V}^i \mathbf{H}_k) \right) - \text{tr}(\mathbf{V}^m \mathbf{H}_k) \leq 0, \quad \forall k \in \mathcal{G}^m, \quad \forall \mathcal{G}^m, \tag{4.28}$$

$$\sum_{m=1}^{M} (1 - c_n^{f^m}) \left| \text{tr}(\mathbf{V}^m \mathbf{J}_n) \right|_0 \log_2 (1 + \Gamma^m) \leq C_{\text{FH},n}, \quad \forall n \in \mathcal{N}, \tag{4.29}$$

$$\sum_{m=1}^{M} \text{tr}(\mathbf{V}^m \mathbf{J}_n) \leq P_{\text{TX},n}^{\max}, \quad \forall n \in \mathcal{N}, \tag{4.30}$$

$$\mathbf{V}^m \succeq \mathbf{0}, \quad \forall m \in \mathcal{M}, \tag{4.31}$$

$$\text{rank}\,(\mathbf{V}^m) = 1, \quad \forall m \in \mathcal{M}. \tag{4.32}$$

For the problem above, it is convex with respect to $\{\mathbf{V}^m\}_{m=1}^M$ except for (4.29) and (4.32). As stated in Subsection 2.3.4, the relaxation step of SDR technique is to drop the rank-one constraint, which is non-convex, and to consider only the remaining relaxed version of the original problem. If the obtained optimal $\mathbf{V}^m$ has rank 1, the EigenValue Decomposition (EVD) can be used to obtain the corresponding optimal beamforming vector $\mathbf{v}^m$. Otherwise randomization and scaling method is used to generate a sub-optimal solution. Details can be found in [KSL08].

After dropping the non-convex constraint (4.32), constraint (4.29) is still non-convex due to the ℓ_0-norm operation inside. In order to convexify it, we utilize the ℓ_0-norm approximation technique introduced in Subsection 2.3.5, i.e., the ℓ_0-norm is approximated in an iterative manner. In each iteration step, a weighted ℓ_1-norm, which is convex, is utilized to approximate the discrete and non-convex ℓ_0-norm, based which a standard Semi Definite Programming (SDP) problem can be generated. By solving the resultant SDP problem, the results are used to recalculate the weights of the ℓ_1-norms so as to refine the approximation. Specifically, in the $(t+1)$-th iteration, $\left|\text{tr}(\mathbf{V}^{m(t+1)} \mathbf{J}_n)\right|_0$ is approximated as a linear function of $\text{tr}(\mathbf{V}^{m(t+1)} \mathbf{J}_n)$, i.e.,

$$\left|\text{tr}(\mathbf{V}^{m(t+1)} \mathbf{J}_n)\right|_0 \approx k_n^{m(t+1)} \text{tr}(\mathbf{V}^{m(t+1)} \mathbf{J}_n), \tag{4.33}$$

where scalar $k_n^{m(t+1)}$ is calculated via the result of the previous iteration as

$$k_n^{m(t+1)} = \frac{1}{\tau + \text{tr}(\mathbf{V}^{m(t)} \mathbf{J}_n)}. \tag{4.34}$$

As said in Subsection 2.3.5, $k_n^{m(t+1)}$ is called the re-weighted coefficient. The value of τ is predetermined and regarded as a threshold parameter that determines whether this ℓ_0-norm is turned on (1) or off (0). Please review Subsection 2.3.5 for more details. In each iteration step, the value of k_n^m shall be updated based on the results of the previous iteration by using (4.34). Then the ℓ_0-norm is approximated for

this iteration via (4.33), which is linear and convex. Therefore, in each iteration step, constraint (4.29) is convexified by such an approximation technique. After the resultant approximated convex problem is solved, we go to the next iteration step with the updated value of k_n^m, then a similar convex problem is formed for the new iteration. As shown in [CWB08], such an iterative approximation of ℓ_0-norm is effective and can converge very fast. The convergence behaviour is also demonstrated in our numerical results, which will be given later.

Therefore, for $(t+1)$-th iteration, the original non-convex problem (4.27)-(4.32) can be relaxed and approximated as

$$\mathcal{P}_{\text{Hard}}^{(t+1)}: \min_{\{\mathbf{V}^{m(t+1)}\}_{m=1}^{M}} \quad \sum_{m=1}^{M} \text{tr}(\mathbf{V}^{m(t+1)}), \tag{4.35}$$

$$\text{s.t.} \quad \Gamma^m \sum_{i\neq m}^{M} \text{tr}(\mathbf{V}^{i(t+1)}\mathbf{H}_k) - \text{tr}(\mathbf{V}^{m(t+1)}\mathbf{H}_k) + \Gamma^m \sigma_k^2 \leq 0, \quad \forall k \in \mathcal{G}^m, \quad \forall \mathcal{G}^m, \tag{4.36}$$

$$\sum_{m=1}^{M} a_n^{m(t+1)} \text{tr}(\mathbf{V}^{m(t+1)}\mathbf{J}_n) - C_{\text{FH},n} \leq 0, \quad \forall n \in \mathcal{N}, \tag{4.37}$$

$$\sum_{m=1}^{M} \text{tr}(\mathbf{V}^{m(t+1)}\mathbf{J}_n) \leq P_{\text{TX},n}^{\max}, \quad \forall n \in \mathcal{N}, \tag{4.38}$$

$$\mathbf{V}^{m(t+1)} \succeq \mathbf{0}, \quad \forall m \in \mathcal{M}, \tag{4.39}$$

where $a_n^{m(t+1)} = k_n^{m(t+1)}(1 - c_n^{f^m}) \log_2(1+\Gamma^m)$, with $k_n^{m(t+1)}$ being calculated according to (4.34), which depends on the results from the previous iteration.

The reformulated problem (4.35)-(4.39) in each iteration consists of only a linear objective function, $K + 2N$ linear inequality constraints, and M positive-semidefinite constraints. It is a standard SDP problem [Fre09] and can be efficiently solved by many solvers, such as SDPT3[TT11] and SeDuMi[PL03].

One important issue is the problem formulation of the initial step, as no previous results exist for the calculation of the value of $k_n^{m(0)}$. The initial value acquisition in this iterative approximation procedure for ℓ_0-norm is circumvented by dropping the constraints that containing the ℓ_0-norm in the initial step. Therefore, for the initial iteration, the following problem shall be solved:

$$\mathcal{P}_{\text{Hard}}^{(0)}: \min_{\{\mathbf{V}^{m(0)}\}_{m=1}^{M}} \quad \sum_{m=1}^{M} \text{tr}(\mathbf{V}^{m(0)}), \tag{4.40}$$

$$\text{s.t.} \quad \Gamma^m \sum_{i\neq m}^{M} \text{tr}(\mathbf{V}^{i(0)}\mathbf{H}_k) - \text{tr}(\mathbf{V}^{m(0)}\mathbf{H}_k) + \Gamma^m \sigma_k^2 \leq 0, \quad \forall k \in \mathcal{G}^m, \quad \forall \mathcal{G}^m, \tag{4.41}$$

$$\sum_{m=1}^{M} \text{tr}(\mathbf{V}^{m(0)}\mathbf{J}_n) \leq P_{\text{TX},n}^{\max}, \quad \forall n \in \mathcal{N}, \tag{4.42}$$

$$\mathbf{V}^{m(0)} \succeq \mathbf{0}, \quad \forall m \in \mathcal{M}. \tag{4.43}$$

By solving the initial problem above, which is also a SDP problem, we can obtain $\{\mathbf{V}^{m(0)}\}_{m=1}^{M}$ which are used to start the iteration steps.

After the last iteration, the final $\{\mathbf{V}^{m(\text{last})}\}_{m=1}^{M}$ can be acquired, and the corresponding beamformers are derived via EigenValue Decomposition (EVD) method or the randomization and scaling method, as introduced in Subsection 2.3.4.

In summary, the overall algorithm for high EE oriented network design with the hard transfer mode for minimizing the transmission power is as follows:

Algorithm 2: The Iterative Optimization Steps for TX power Minimization (*For the hard transfer mode*)

Initialization: Solve the standard SDP problem $\mathcal{P}_{\text{Hard}}^{(0)}$ (4.40)-(4.43) to obtain $\{\mathbf{V}^{m(0)}\}_{m=1}^{M}$. Compute $k_n^{m(1)}$ based on (4.34), $\forall m, n$. Construct the problem $\mathcal{P}_{\text{Hard}}^{(1)}$ according to (4.35)-(4.39), and set $t \leftarrow 1$.
repeat
  Solve the standard SDP problem $\mathcal{P}_{\text{Hard}}^{(t)}$ for obtaining $\{\mathbf{V}^{m(t)}\}_{m=1}^{M}$.
  Update the values of $k_n^{m(t+1)}$ based on (4.34), $\forall m, n$. Then formulate the problem $\mathcal{P}_{\text{Hard}}^{(t+1)}$ according to (4.35)-(4.39), and set $t \leftarrow t+1$.
until *convergence or reaching the max iteration number*;
if $\text{rank}(\mathbf{V}^{m(\text{last})}) = 1$ **then**
  Perform EVD to obtain the optimal $\{\mathbf{v}^{m}\}_{m=1}^{M}$.
else
  Use Gaussian randomization and scaling [KSL08] method to obtain the approximate solution $\{\mathbf{v}^{m}\}_{m=1}^{M}$.

Extension to the non-dedicated case: When the fronthaul capacity is shared among eRRHs, we can still use the same techniques proposed above to solve the problem. Comparing the inequality between (4.24) and (4.23), they have almost the same formulation. Hence, we can adopt the same iterative ℓ_0-norm approximation method to convexify the capacity constraint of the non-dedicated fronthaul. The algorithm above is still valid. After the final optimized beamforming vectors $\{\mathbf{v}^{m,\text{opt}}\}_{m=1}^{M}$ are obtained via the algorithm, the fronthaul capacity $C_{\text{FH},n}^{\text{hard,opt}}$ that shall be allocated to eRRH n can be derived by calculating

$$C_{\text{FH},n}^{\text{hard,opt}} = \sum_{m=1}^{M} (1 - c_n^{f^m}) H\left(||\mathbf{v}_n^{m,\text{opt}}||_2^2, \tau\right) \log_2\left(1 + \Gamma^m\right), \tag{4.44}$$

where the unit step function $H(x, \tau)$ used here is defined as

$$H(x, \tau) := \begin{cases} 0 & \text{for } x \leq \tau \\ 1 & \text{for } x > \tau \end{cases} \tag{4.45}$$

τ is the predetermined threshold parameter that has been used for the ℓ_0-norm approximation.

4.2.1.2 Design for the Soft Transfer Mode

For soft transfer mode, we formulate the problem to be solved according to (4.9), (4.10), (4.14) and (4.16), as follows:

$$\mathcal{P}_{\text{Soft original}}: \quad \min_{\{\mathbf{w}^m\}_{m=1}^M, \{\mathbf{q}_n\}_{n=1}^N} \left(\sum_{m=1}^{M} ||\mathbf{w}^m||_2^2 + \sum_{n=1}^{N} ||\mathbf{q}_n||_2^2 \right), \tag{4.46}$$

$$\text{s.t.} \quad \text{SINR}_k^{\text{soft}} \geq \Gamma^m, \quad \forall k \in \mathcal{G}^m, \quad \forall \mathcal{G}^m, \tag{4.47}$$

$$\sum_{l=1}^{L} \log_2 \left(1 + \frac{\sum_{m=1}^{M} (1 - c_n^{f^m}) |w_{n,l}^m|^2}{q_{n,l}^2} \right) \leq C_{\text{FH},n} \quad \forall n \in \mathcal{N} \quad \textit{dedicated}, \tag{4.48}$$

$$\sum_{n=1}^{N} \sum_{l=1}^{L} \log_2 \left(1 + \frac{\sum_{m=1}^{M} (1 - c_n^{f^m}) |w_{n,l}^m|^2}{q_{n,l}^2} \right) \leq C_{\text{FH}} \quad \textit{non-dedicated}, \tag{4.49}$$

$$\sum_{m=1}^{M} ||\mathbf{w}_n^m||_2^2 + ||\mathbf{q}_n||_2^2 \leq P_{\text{TX},n}^{\max}, \quad \forall n \in \mathcal{N}. \tag{4.50}$$

Due to the inevitable quantization error introduced by the compression procedure in the soft transfer mode, each eRRH has to reserve some power for the transmission of the quantization noise. Hence, when minimizing the total transmission power in (4.46), besides optimizing the precoders, the optimal design of the introduced quantization noise shall also be taken into account.

Similarly to the hard transfer mode, constraint (4.47) guarantees the QoS of each UE in each multi-cast group, where Γ^m denotes the target SINR of the content requested by multi-cast group $\mathcal{G}^m$. $\text{SINR}_k^{\text{soft}}$ has been derived in (4.16). Constraints (4.48) and (4.49) guarantee that fronthaul can support the soft transfer of the data streams to eRRHs, for the dedicated and the non-dedicated scenarios, respectively. Constraint (4.50) ensures that the transmission power of each eRRH does not exceed its maximal allowable power.

The solution of the problem above will not only give the optimal precoder design, but also the optimal compression strategy for the soft transferring of data streams to each eRRHs from the BBU pool. Obviously, this problem is also non-convex. Similar to the solving strategy of the hard transfer mode, we relax, reformulate, approximate and then convexify the original problem to make it solvable. Similarly, we take the **dedicated** case as the example to illustrate the solving procedure, i.e., (4.46)-(4.48) and (4.50). The extension to the non-dedicated case will be given afterwards.

By comparing the problem formulation of the soft transfer mode, with the one for the hard transfer mode (4.21)-(4.25), we see that the SDR technique can be adopted to convexify (4.46), (4.47) and (4.50). The problem becomes how to convexify (4.48).

We adopt an iterative approximation method for the convexification of it, whose convergence and effectiveness are proved and shown in [DW16]. Specifically, by adopting the SDR technique, the Left Hand Side (LHS) of (4.48) can be reformulated as follows and be upper-bounded:

$$
\begin{aligned}
&\sum_{l=1}^{L}\log_2\left(1+\frac{\sum_{m=1}^{M}(1-c_n^{f^m})|w_{n,l}^m|^2}{q_{n,l}^2}\right)\\
&=\sum_{l=1}^{L}\log_2\left(\frac{q_{n,l}^2+\sum_{m=1}^{M}(1-c_n^{f^m})|w_{n,l}^m|^2}{q_{n,l}^2}\right)\\
&=\sum_{l=1}^{L}\log_2\left(\frac{\mathrm{tr}\left(\mathbf{Q}\mathbf{J}_{n,l}\right)+\sum_{m=1}^{M}(1-c_n^{f^m})\mathrm{tr}\left(\mathbf{W}^m\mathbf{J}_{n,l}\right)}{\mathrm{tr}\left(\mathbf{Q}\mathbf{J}_{n,l}\right)}\right)\\
&=\sum_{l=1}^{L}\log_2\left(\mathrm{tr}\left(\mathbf{Q}\mathbf{J}_{n,l}\right)+\sum_{m=1}^{M}(1-c_n^{f^m})\mathrm{tr}\left(\mathbf{W}^m\mathbf{J}_{n,l}\right)\right)-\sum_{l=1}^{L}\log_2\mathrm{tr}\left(\mathbf{Q}\mathbf{J}_{n,l}\right)\\
&\le\sum_{l=1}^{L}\left(\log_2\eta_{n,l}+\frac{\mathrm{tr}\left(\mathbf{Q}\mathbf{J}_{n,l}\right)+\sum_{m=1}^{M}(1-c_n^{f^m})\mathrm{tr}\left(\mathbf{W}^m\mathbf{J}_{n,l}\right)}{\eta_{n,l}\ln 2}\right)\\
&\quad-\frac{L}{\ln 2}-\sum_{l=1}^{L}\log_2\mathrm{tr}\left(\mathbf{Q}\mathbf{J}_{n,l}\right), \qquad (4.51)\\
&\mathrm{rank}\left(\mathbf{W}^m\right)=1,\quad \forall m\in\mathcal{M}, \qquad (4.52)\\
&\mathrm{rank}\left(\mathbf{Q}\right)=1, \qquad (4.53)
\end{aligned}
$$

where $\mathbf{W}^m=\mathbf{w}^m\mathbf{w}^{mH}\in\mathbb{R}^{NL\times NL}\ \forall m\in\mathcal{M}$ and $\mathbf{Q}=\mathbf{q}\mathbf{q}^H\in\mathbb{R}^{NL\times NL}$ are positive semidefinite matrices, i.e., $\mathbf{Q},\ \mathbf{W}^m\succeq\mathbf{0}$. The antenna selection matrix $\mathbf{J}_{n,l}\in\mathbb{R}^{NL\times NL}$ is a diagonal matrix, whose $((n-1)L+l)$-th diagonal element is 1, all others are 0.

The main point lies in how can we obtain the inequality in (4.51). Note that according to Bernoulli's Inequality, for any $x\in\mathbb{R}$, we have

$$1+x\le e^x, \qquad (4.54)$$

thus,

$$\ln x\le x-1. \qquad (4.55)$$

We achieve the equality in (4.55) when $x=1$.

By introducing an auxiliary parameter $\eta_{n,l}$ to the LHS of (4.51) and adopting (4.55),

we have

$$\ln \frac{\mathrm{tr}\left(\mathbf{Q}\mathbf{J}_{n,l}\right)+\sum_{m=1}^{M}(1-c_n^{f^m})\mathrm{tr}\left(\mathbf{W}^m\mathbf{J}_{n,l}\right)}{\eta_{n,l}} \leq \frac{\mathrm{tr}\left(\mathbf{Q}\mathbf{J}_{n,l}\right)+\sum_{m=1}^{M}(1-c_n^{f^m})\mathrm{tr}\left(\mathbf{W}^m\mathbf{J}_{n,l}\right)}{\eta_{n,l}}-1, \tag{4.56}$$

then,

$$\frac{1}{\ln 2}\ln \frac{\mathrm{tr}\left(\mathbf{Q}\mathbf{J}_{n,l}\right)+\sum_{m=1}^{M}(1-c_n^{f^m})\mathrm{tr}\left(\mathbf{W}^m\mathbf{J}_{n,l}\right)}{\eta_{n,l}} \leq \frac{1}{\ln 2}\frac{\mathrm{tr}\left(\mathbf{Q}\mathbf{J}_{n,l}\right)+\sum_{m=1}^{M}(1-c_n^{f^m})\mathrm{tr}\left(\mathbf{W}^m\mathbf{J}_{n,l}\right)-\eta_{n,l}}{\eta_{n,l}}. \tag{4.57}$$

By adopting the conversion

$$\log_a b = \frac{\ln b}{\ln a},$$

we further have

$$\begin{aligned}&\log_2\left(\mathrm{tr}\left(\mathbf{Q}\mathbf{J}_{n,l}\right)+\sum_{m=1}^{M}(1-c_n^{f^m})\mathrm{tr}\left(\mathbf{W}^m\mathbf{J}_{n,l}\right)\right)-\log_2\eta_{n,l}\\ &\leq \frac{1}{\ln 2}\frac{\mathrm{tr}\left(\mathbf{Q}\mathbf{J}_{n,l}\right)+\sum_{m=1}^{M}(1-c_n^{f^m})\mathrm{tr}\left(\mathbf{W}^m\mathbf{J}_{n,l}\right)-\eta_{n,l}}{\eta_{n,l}},\end{aligned} \tag{4.58}$$

and finally

$$\begin{aligned}&\log_2\left(\mathrm{tr}\left(\mathbf{Q}\mathbf{J}_{n,l}\right)+\sum_{m=1}^{M}(1-c_n^{f^m})\mathrm{tr}\left(\mathbf{W}^m\mathbf{J}_{n,l}\right)\right)\\ &\leq \log_2\eta_{n,l}+\frac{\mathrm{tr}\left(\mathbf{Q}\mathbf{J}_{n,l}\right)+\sum_{m=1}^{M}(1-c_n^{f^m})\mathrm{tr}\left(\mathbf{W}^m\mathbf{J}_{n,l}\right)}{\eta_{n,l}\ln 2}-\frac{1}{\ln 2},\end{aligned} \tag{4.59}$$

the equality in (4.59) holds if and only if

$$\eta_{n,l}=\mathrm{tr}\left(\mathbf{Q}\mathbf{J}_{n,l}\right)+\sum_{m=1}^{M}(1-c_n^{f^m})\mathrm{tr}\left(\mathbf{W}^m\mathbf{J}_{n,l}\right), \quad \forall l\in\{1,2,...,L\}. \tag{4.60}$$

Hence, we can finally upper bound the LHS of (4.51) and obtain the corresponding inequality.

For fixed values of $\{\eta_{n,l}\}_{l=1}^{L}\ \forall n\in\mathcal{N}$, the Right Hand Side (RHS) of (4.51) is a convex function with respect to $\mathbf{Q}$ and $\mathbf{W}^m$, which motivates a successive solving strategy of the original problem: By replacing the LHS of (4.48) with the RHS of (4.51) for specific $\{\eta_{n,l}\}_{l=1}^{L}\forall n\in\mathcal{N}$, whose values are obtained from the results of the problem from the previous iteration. Then in each iteration, a relaxed convex optimization problem can be formulated, by temporarily dropping the non-convex constraint (4.52) and (4.53).

Specifically, after such a convex problem is solved in each iteration, the values of $\{\eta_{n,l}\}_{l=1}^{L}\ \forall n\in\mathcal{N}$ are to be updated according to (4.60), which are then utilized to

formulate the problem of the next iteration. This technique is very similar to how we dealt with the non-convex ℓ_0-norm previously.

Similarly, the final $\{\mathbf{w}^m\}_{m=1}^M$ and $\mathbf{q}$ can be derived via EVD or Gaussian randomization and scaling, as introduced several times in previous subsections.

In detail, by combining the SDR technique and the iterative approximation technique proposed above, for the $(t+1)$-th iteration, the problem is formulated as follows:

$$\mathcal{P}_{\text{Soft}}^{(t+1)}: \min_{\{\mathbf{W}^{m(t+1)}\}_{m=1}^M, \mathbf{Q}^{(t+1)}} \quad \sum_{m=1}^{M} \text{tr}\left(\mathbf{W}^{m(t+1)}\right) + \sum_{n=1}^{N} \text{tr}\left(\mathbf{Q}^{(t+1)}\mathbf{J}_n\right), \tag{4.61}$$

$$\text{s.t.} \quad \Gamma^m \text{tr}(\mathbf{Q}^{(t+1)}\mathbf{H}_k) + \Gamma^m \sum_{i\neq m}^{M} \text{tr}(\mathbf{W}^{i(t+1)}\mathbf{H}_k) - \text{tr}(\mathbf{W}^{m(t+1)}\mathbf{H}_k) + \Gamma^m \sigma_k^2 \leq 0, \quad \forall k \in \mathcal{G}^m, \quad \forall \mathcal{G}^m, \tag{4.62}$$

$$\sum_{l=1}^{L}\left(\log_2 \eta_{n,l} + \frac{\text{tr}\left(\mathbf{Q}^{(t+1)}\mathbf{J}_{n,l}\right) + \sum_{m=1}^{M}(1-c_n^{f^m})\text{tr}\left(\mathbf{W}^{m(t+1)}\mathbf{J}_{n,l}\right)}{\eta_{n,l}\ln 2}\right) - \sum_{l=1}^{L}\log_2 \text{tr}\left(\mathbf{Q}^{(t+1)}\mathbf{J}_{n,l}\right) - \frac{L}{\ln 2} - C_{\text{FH},n} \leq 0, \quad \forall n \in \mathcal{N}, \tag{4.63}$$

$$\sum_{m=1}^{M} \text{tr}(\mathbf{W}^{m(t+1)}\mathbf{J}_n) + \text{tr}(\mathbf{Q}^{(t+1)}\mathbf{J}_n) \leq P_{\text{TX},n}^{\max}, \quad \forall n \in \mathcal{N}, \tag{4.64}$$

$$\mathbf{W}^{m(t+1)} \succeq \mathbf{0}, \quad \forall m \in \mathcal{M}, \tag{4.65}$$

$$\mathbf{Q}^{(t+1)} \succeq \mathbf{0}. \tag{4.66}$$

The reformulated problem (4.61)-(4.66) in each iteration consists of a linear objective function, $K+2N$ linear inequality constraints, and $M+1$ positive-semidefinite constraints. Hence, it is also a SDP problem and can be solved by SDPT3 or SeDuMi introduced previously.

Similarly, an initial step is required to compute values of $\{\eta_{n,l}\}_{l=1}^L \; \forall n \in \mathcal{N}$ that are to be used in next iterations. Same as the initial step for solving the problem of the hard transfer mode, the constraints where $\{\eta_{n,l}\}_{l=1}^L \; \forall n \in \mathcal{N}$ appear are temporarily dropped, i.e., the initial problem of iteration 0 is formed as follows without the

fronthaul constraints (4.63):

$$\mathcal{P}^{(0)}_{\text{Soft}}: \min_{\{\mathbf{W}^{m(0)}\}_{m=1}^{M},\mathbf{Q}^{(0)}} \quad \sum_{m=1}^{M} \text{tr}\left(\mathbf{W}^{m(0)}\right) + \sum_{n=1}^{N} \text{tr}\left(\mathbf{Q}^{(0)}\mathbf{J}_n\right), \tag{4.67}$$

$$\text{s.t.} \quad \Gamma^m \text{tr}(\mathbf{Q}^{(0)}\mathbf{H}_k) + \Gamma^m \sum_{i\neq m}^{M} \text{tr}(\mathbf{W}^{i(0)}\mathbf{H}_k) - \text{tr}(\mathbf{W}^{m(0)}\mathbf{H}_k) + \Gamma^m\sigma_k^2 \leq 0,$$

$$\forall k \in \mathcal{G}^m, \quad \forall \mathcal{G}^m, \tag{4.68}$$

$$\sum_{m=1}^{M} \text{tr}(\mathbf{W}^{m(0)}\mathbf{J}_n) + \text{tr}(\mathbf{Q}^{(0)}\mathbf{J}_n) \leq P^{\max}_{\text{TX},n}, \quad \forall n \in \mathcal{N}, \tag{4.69}$$

$$\mathbf{W}^{m(0)} \succeq \mathbf{0}, \quad \forall m \in \mathcal{M}, \tag{4.70}$$

$$\mathbf{Q}^{(0)} \succeq \mathbf{0}. \tag{4.71}$$

After solving the initial SDP problem above, all required initial values to compute $\{\eta_{n,l}\}_{l=1}^{L}$ $\forall n \in \mathcal{N}$ for further iterations can be obtained according to (4.60). The overall algorithm is summarized as follows:

Algorithm 3: The Iterative Optimization Steps for TX power Minimization (*For the soft transfer mode*)

1 **Initialization**: Solve the standard SDP problem $\mathcal{P}^{(0)}_{\text{Soft}}$ (4.67)-(4.71) to obtain $\{\mathbf{W}^{m(0)}\}_{m=1}^{M}$ and $\mathbf{Q}^{(0)}$. Compute $\eta^{(1)}_{n,l}$ based on (4.60), $\forall n, l$. Construct the problem $\mathcal{P}^{(1)}_{\text{Soft}}$ according to (4.61)-(4.66), and set $t \leftarrow 1$.
2 **repeat**
3 Solve the standard SDP problem $\mathcal{P}^{(t)}_{\text{Soft}}$ for obtaining $\{\mathbf{V}^{m(t)}\}_{m=1}^{M}$ and $\mathbf{Q}^{(t)}$.
4 Compute the values of $\eta^{(t+1)}_{n,l}$ based on (4.60), $\forall n, l$. Then formulate the problem $\mathcal{P}^{(t+1)}_{\text{Soft}}$ according to (4.61)-(4.66), and set $t \leftarrow t+1$.
5 **until** *convergence or reaching the max iteration number*;
6 **if** $\text{rank}(\mathbf{W}^{m(\text{last})}) = 1$ *and* $\text{rank}(\mathbf{Q}^{(\text{last})}) = 1$ **then**
7 Perform EVD to obtain the optimal $\{\mathbf{w}^m\}_{m=1}^{M}$ and $\mathbf{q}$.
8 **else**
9 Use Gaussian randomization and scaling [KSL08] method to obtain the approximate solution $\{\mathbf{w}^m\}_{m=1}^{M}$ and $\mathbf{q}$.

With the obtained $\mathbf{q}$ from the algorithm, the BBU pool acquires the statistical knowledge for the quantization step of the soft transfer mode, more details can be backtracked to Subsection 4.2.1.2.

Extension to the non-dedicated case: When the fronthaul capacity is shared among eRRHs, by comparing the constraints (4.48) and (4.49), it can be concluded that the technique introduced above can still be adopted to solve the problem. Thus, the same iterative approximation method can be utilized to convexify the capacity constraint of the non-dedicated fronthaul (4.49), and Alg. 3 is thus still valid. After

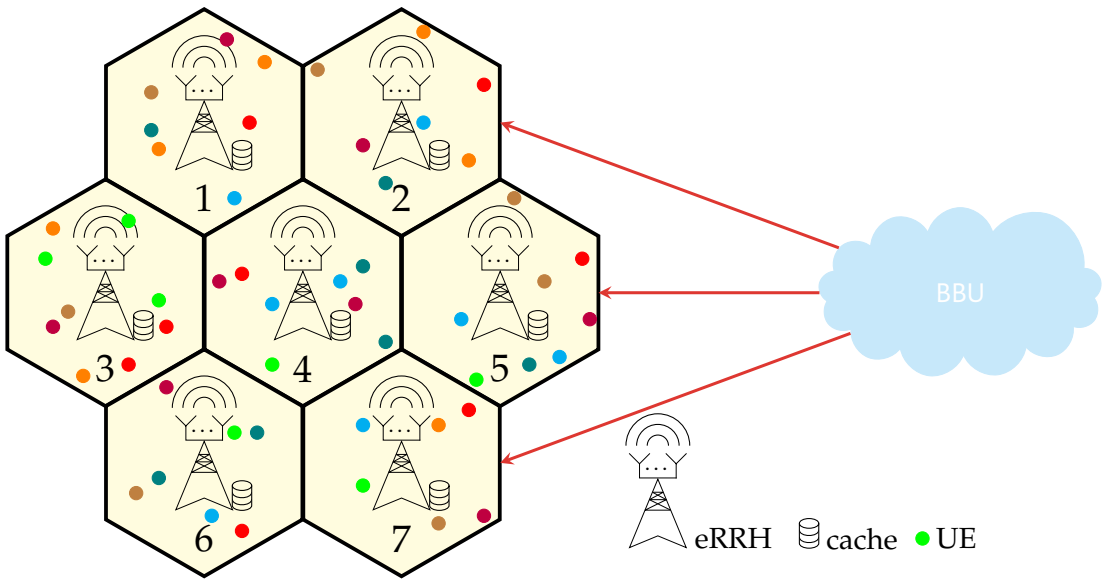


Figure 4.4: The cache-enabled F-RAN consisting of seven hexagonal cells used for simulation in Subsection 4.2.1.3 and several later subsections. Dots with the same color denote UEs requesting the same contents, which are randomly and uniformly distributed within the whole network. The index for each eRRH/cell lies at the bottom of each hexagon.

the final optimized $\{\mathbf{w}^{m,\text{opt}}\}_{m=1}^{M}$ and $\mathbf{q}^{\text{opt}}$ are obtained by solving the algorithm, the fronthaul capacity $C_{\text{FH},n}^{\text{soft,opt}}$ that shall be allocated to eRRH n can be derived by

$$C_{\text{FH},n}^{\text{soft,opt}} = \sum_{l=1}^{L} \log_2 \left(1 + \frac{\sum_{m=1}^{M} (1 - c_n^{f^m}) |w_{n,l}^{m,\text{opt}}|^2}{(q_{n,l}^{\text{opt}})^2} \right). \tag{4.72}$$

4.2.1.3 Numerical Results

It is time to have a pause here before we continue our algorithm adventure. The numerical results of the two algorithms proposed above will be provided in this subsection to verify their correctness and effectiveness. Furthermore, the performance of the hard and soft transfer modes will also be compared. A hexagonal F-RAN is selected as illustrated in Fig.4.4, the wireless environment is setup with the parameters listed in Table 4.1 [4], all simulation results are based on these parameters unless otherwise stated. We adopt the system model, including the network model, cache and content model, etc., according to the descriptions in Section 4.1. All UEs are randomly and uniformly distributed within this hexagonal network.

Our simulation results are to be compared with some existing algorithms in other works, with which the benefit and effectiveness of the proposed ones can be demonstrated.

[4] For the fronthaul capacity $C_{\text{FH},n}$, we assume the dedicated capacity for the pilot signals have been deducted.

Number of eRRHs (Hexagonal Cell): N	7
Number of antennas per eRRH: L	2
Distance between adjacent eRRHs: d_{eRRH}	0.5 km
Transmit Antenna Gain	10 dBi
Total number of UE: K_{total}	200
Number of scheduled UEs per DL slot: K	12
Background noise	-172 dBm/Hz
3GPP LTE-A path loss model	$148.1 + 37.6 \log_{10}(d)$
Log-normal shadowing	8 dB
Rayleigh small scale fading	0 dB
Network bandwidth: B	10 MHz
SINR target for each UE: Γ	10 dB
Total number of contents: M_{total}	100
Skew parameter of the Zipf distribution: α	1.5
Cache Memory Size: S	3 Units
Individual fronthaul capacities: $C_{\text{FH},n}$ $\forall n \in \{1, 2, ..., N\}$	70Mbps
Threshold parameter used in (4.34): τ	-50dBm
Maximal iteration number: $N_{\max}$:	30

Table 4.1: The simulation parameters for F-RAN.

Firstly, we test the proposed algorithm for the hard transfer mode, i.e., Alg. 2. In the specific downlink slot selected in our simulation, according to the network configuration, **twelve** scheduled UEs are allowed to submit their requests. The BBU pool realizes that totally **seven** different contents are requested, thus seven multi-cast groups are formed. Among seven requested contents, two of them have already been cached at eRRHs, whose indices are 1 and 2, respectively. Although the cache memory size S is 3, the cached content with index 3 is not requested by any UE in this downlink slot. Without loss of generality, we name the two requested contents that are cached as $f(1)$ and $f(2)$ for multi-cast group 1 and 2, respectively. The remaining five requested contents, which are named by $f(3)$ to $f(7)$, have to be fetched from the cloud via fronthauls. After the BBU pool knows such knowledge, Alg. 2 is executed to optimize the network. As a comparison to our proposed algorithm, the algorithm proposed in [Tao+16] is also implemented. Moreover, we record the value of $P_n^m = ||\mathbf{v}_n^m||_2^2$ in each iteration. It denotes how much power is allocated for transmitting content $f(m)$ at eRRH n, in order to serve the UEs in multi-cast group $\mathcal{G}^m$. Moreover, the optimal clustering pattern, i.e., which subset of eRRHs shall serve which multi-cast group, can be derived.

In Fig. 4.5 - Fig. 4.8, the clustering patterns are illustrated from the perspective of the requested contents. In Fig. 4.5 and Fig. 4.6, the y-axis denotes the allocated power for the cached file $f(2)$[C] at all seven eRRHs, which are plotted with solid lines. The x-axis indicates the iteration number of the running algorithms. Fig. 4.7 and Fig. 4.8 illustrate the allocated power for the uncached content $f(6)$[U], which are plotted as dotted-dashed lines. For both contents, the results are acquired by executing the

Figure 4.5: The cluster for cached content $f(2)$[C] resulting from Alg. 2.

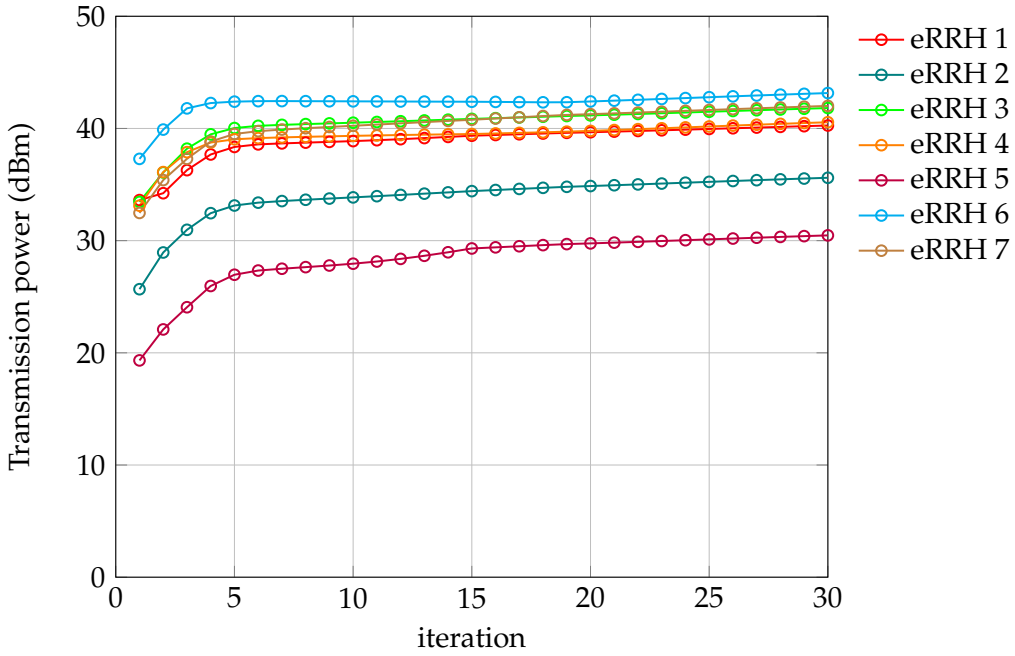


Figure 4.6: The cluster for cached content $f(2)$[C] resulting from the benchmark Alg. in [Tao+16].

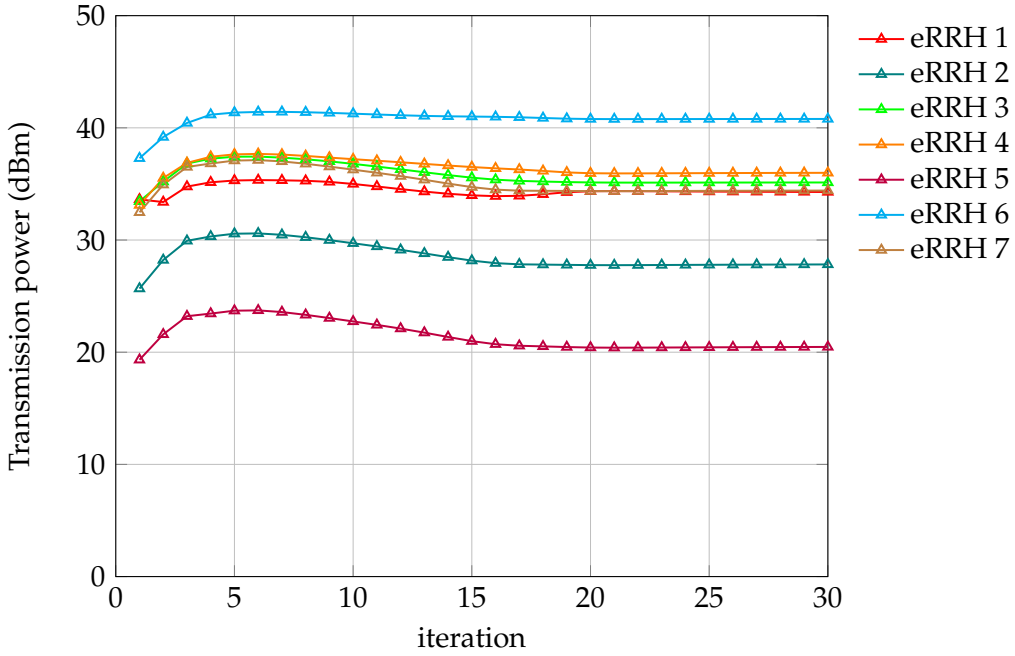


Figure 4.7: The cluster for uncached content $f(6)$[U] resulting from Alg. 2.

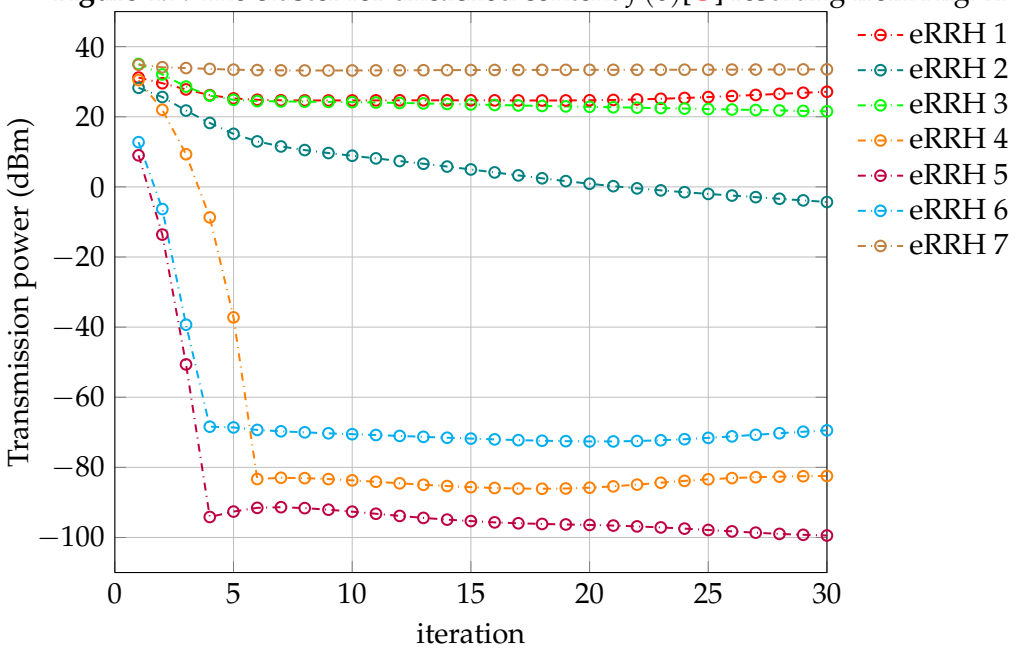
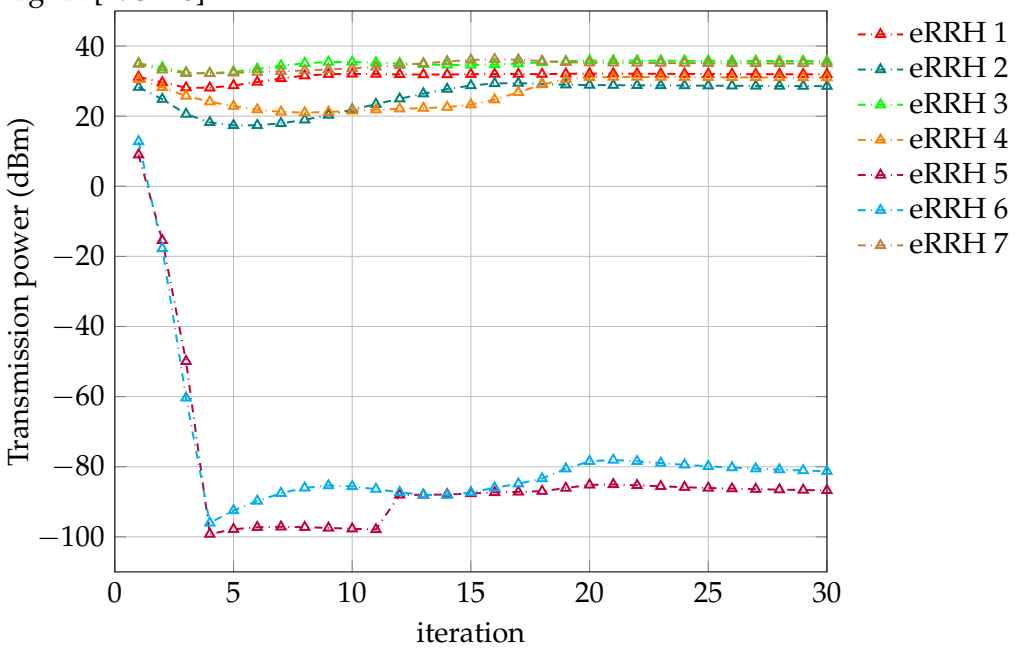


Figure 4.8: The cluster for uncached content $f(6)$[U] resulting from the benchmark Alg. in [Tao+16].

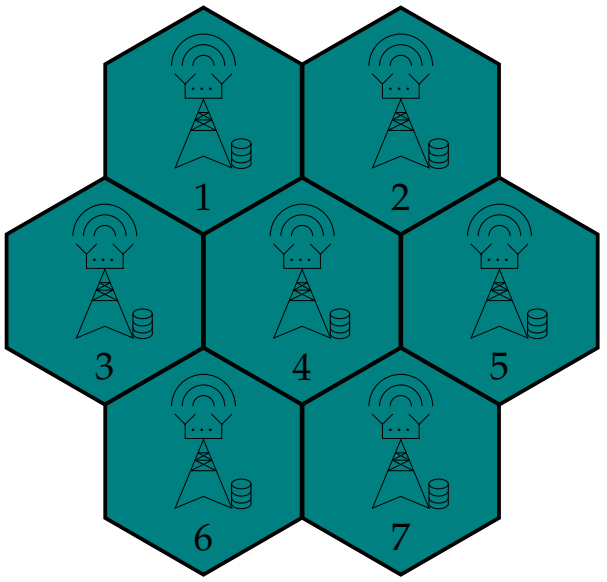


proposed Alg. 2, and the benchmark algorithm from [Tao+16], respectively.

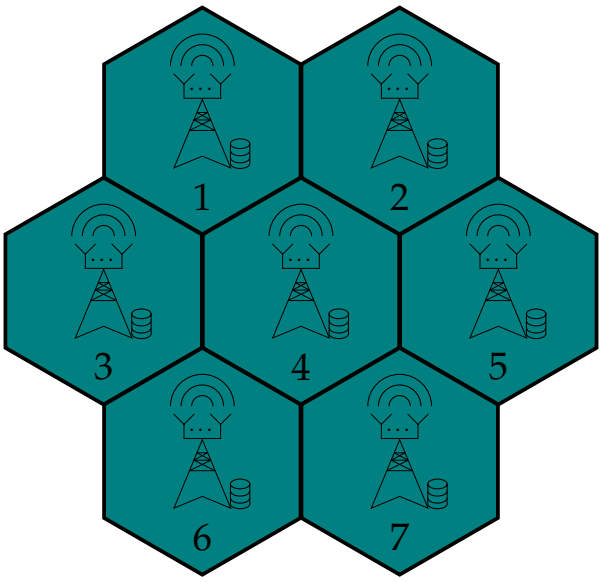
Note that the threshold parameter τ is set to -50 dBm. Hence, for cached content $f(2)$[C], we see that all seven eRRHs from the results of both algorithms shall participate in transmitting this content, as no one drops below -50 dBm. This is just what we expected: As the cached contents do not consume fronthaul resources, involving all eRRHs in this cluster can always increase the spatial diversity and thus reduce the transmission power consumption. Hence, for the cached contents, the clustering results are the same for both algorithm. Additionally, we see that the proposed algorithm converges and yields stable outcomes just after about five iterations. However, when it goes to uncached content, involving all eRRHs in a cluster for a single content might be not possible any more: As the fronthaul resources are consumed, delivering the uncached contents to all eRRHs so as to increase the spatial diversity might not be supported. Hence, for each uncached content, some eRRHs shall be expelled out of the participation for the transmission of them. We take uncached content $f(6)$[U] as an example: By comparing Fig. 4.7 and Fig. 4.8, it can be observed that the resultant clustering pattern are different: In Fig. 4.7, in order to meet each individual fronthaul capacity constraint, our proposed algorithm expels three eRRHs, i.e., eRRH 4, eRRH 5 and eRRH 6, out of the cluster for transmitting content $f(6)$[U], after about six iterations. However, with the benchmark algorithm, only two eRRHs, i.e., eRRH 5 and eRRH 6, are expelled from the eRRH-cluster to serve $f(6)$[U]. In order to illustrate these results more intuitive and easier to understand, we plot the final cluster formulation in Fig. 4.9, which is resultant from the outcome of Fig. 4.5 - Fig. 4.8. Obviously, for uncached content $f(6)$[U], the eRRH-cluster formed via these two algorithms are different, as we are going to show next, the cluster formed via the benchmark algorithm actually causes traffic problems.

For demonstrating how the proposed algorithm regulates the traffic on fronthauls, Fig. 4.10 - Fig. 4.13 are plotted. These figures are obtained with the same simulation realization as Fig. 4.5 - Fig. 4.8, however, the cluster formulation is plotted from the perspective of eRRHs. As stated before, each eRRH might participate in several clusters for serving different multi-cast groups. Note that the fronthaul capacity of each eRRH is set to be 70 Mbps, thus, besides supporting two cached contents without consuming the fronthaul resources, each eRRH can support at most two uncached data streams, via a simple computation: $B\log_2(1+\Gamma)\times 2 = 10\times\log_2(1+10)\approx 70$ Mbps. In Fig. 4.10 and Fig. 4.12, it can be observed that with the proposed algorithm, the cluster is formulated such that exactly two data streams of the uncached contents are transmitted by eRRH 3 and 5, i.e., $f(5)$[U], $f(6)$[U] are supported by eRRH 3, and $f(4)$[U], $f(5)$[U] are supported by eRRH 5. They all participate in transmitting two cached contents, plus additional two uncached contents. However, with the algorithm proposed in [Tao+16], as shown in Fig. 4.11 and

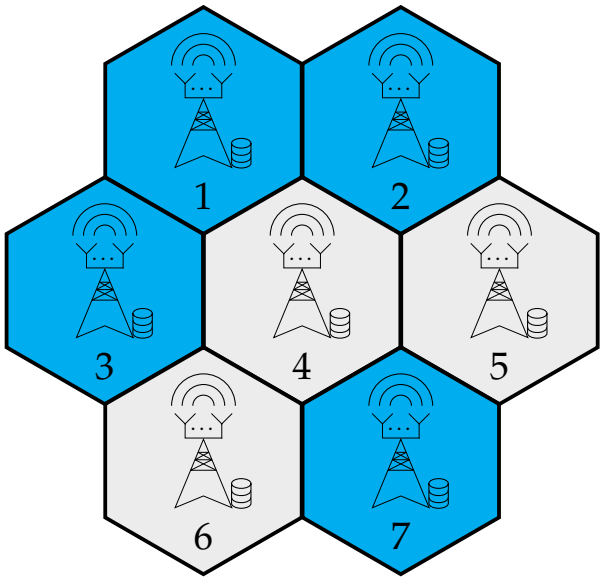
(a) Cluster for cached $f(2)$ (Proposed)



(b) Cluster for cached $f(2)$ (Benchmark)



(c) Cluster for uncached $f(6)$ (Proposed)



(d) Cluster for uncached $f(6)$ (Benchmark)

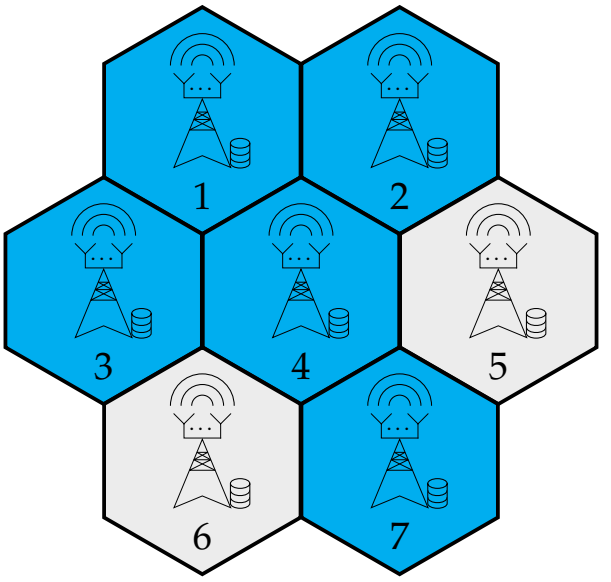


Figure 4.9: An illustration of the final cluster formulation for $f(2)$[C] and $f(6)$[U] with the proposed Alg. 2 and the benchmark Alg. in [Tao+16]. Colored cell denotes that the eRRH mounted within this cell is determined to be in the cluster to serve the corresponding content/multi-cast group. Cells colored with light gray indicates that the eRRH mounted within this cell shall not be involved in this cluster.

Figure 4.10: The cluster involvement of eRRH 3 for all contents resulting from Alg. 2.

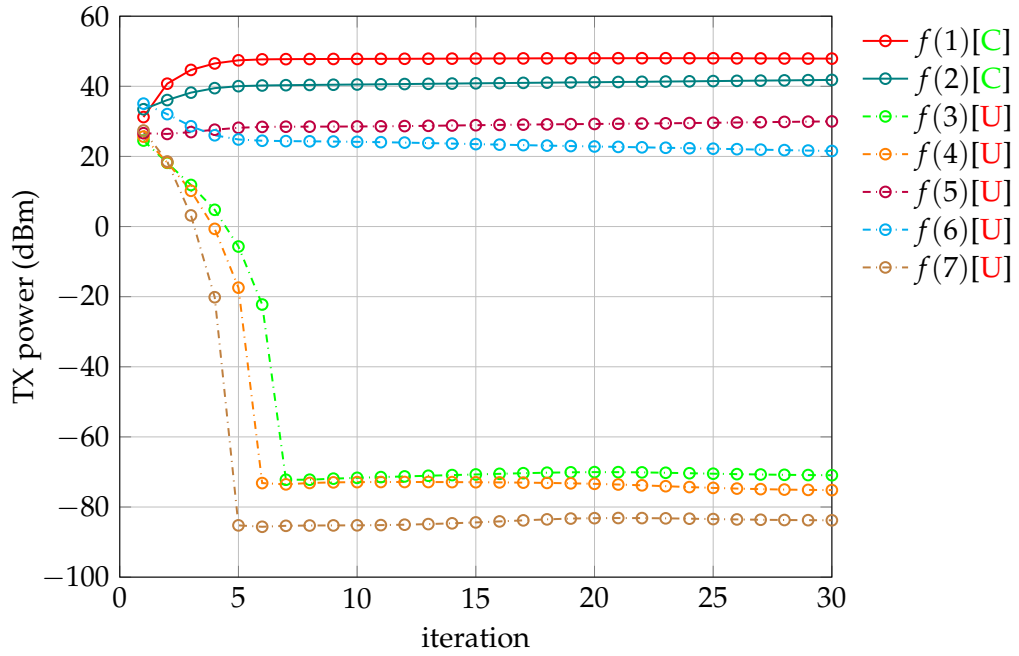


Figure 4.11: The cluster involvement of eRRH 3 for all contents resulting from the benchmark Alg. in [Tao+16].

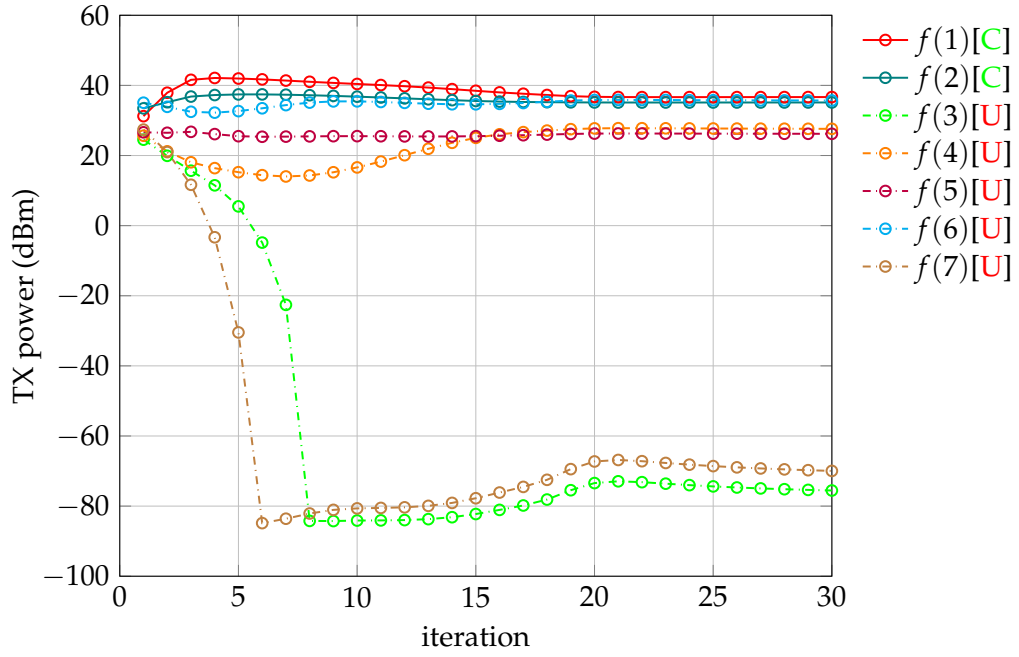


Figure 4.12: The cluster involvement of eRRH 5 for all contents resulting from Alg. 2.

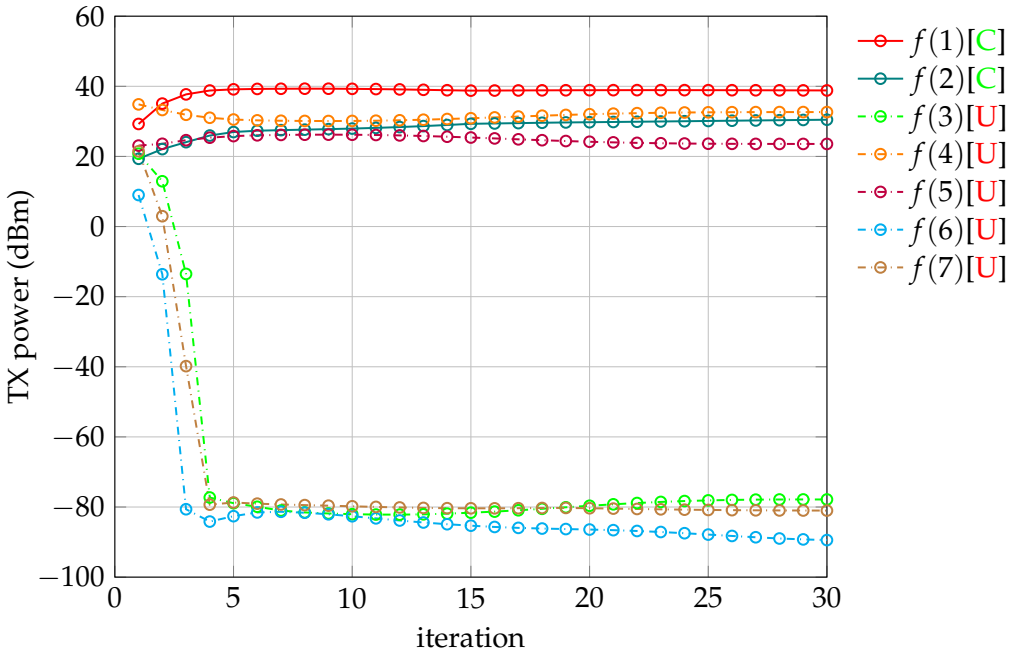


Figure 4.13: The cluster involvement of eRRH 5 for all contents resulting from the benchmark Alg. in [Tao+16].

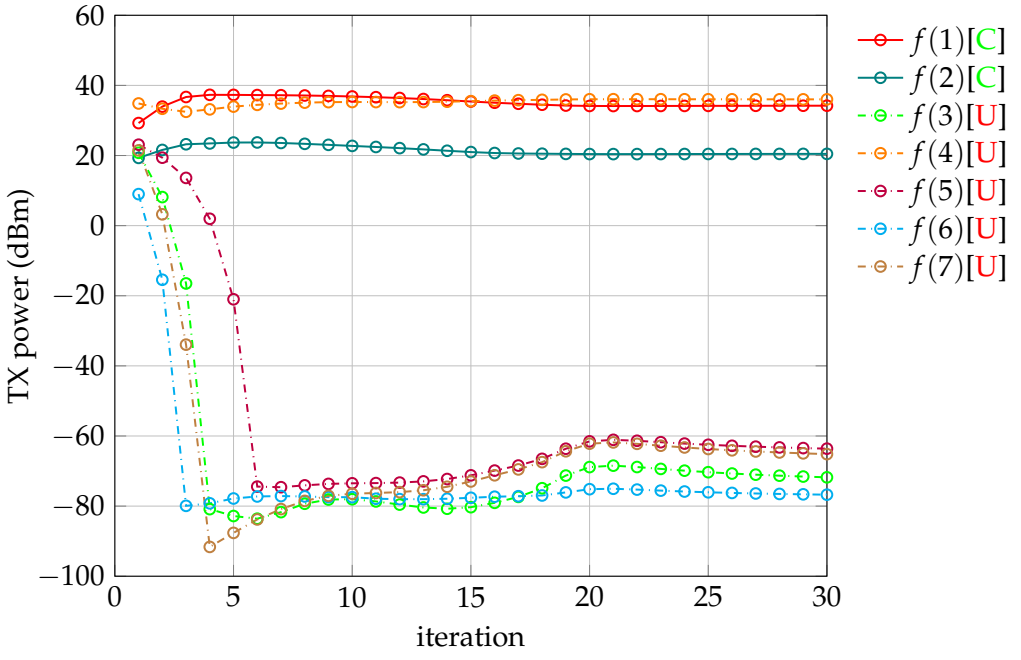


Fig. 4.13, eRRH 3 has to support three data streams of the uncached contents, i.e., $f(4)$[U], $f(5)$[U] and $f(6)$[U], but eRRH 5 supports only one uncached data stream, i.e., $f(4)$[U]. Hence, the results obtained by [Tao+16] cause traffic congestion, e.g., at eRRH 3, and resource waste, e.g., at eRRH 5. Similarly, to be more intuitive and for easier understanding, we plot the final clustering results of this two eRRHs for this specific slot in Fig. 4.14, where the results from both algorithms are presented.

Now we have shown the iterative behaviour of the proposed algorithm, and how clusters are formed to satisfy the fronthaul capacity constraints. Next, let's look at the resultant minimized transmission power, the results are shown in Fig. 4.15.

In Fig. 4.15, the total transmission power consumption for different individual fronthaul capacities [5] and different number of requested contents that have been cached, are compared. Obviously, the results demonstrate that the transmission power consumption can be reduced either by caching more contents, or by increasing the fronthaul capacity, due to more cooperation among eRRHs becoming possible. However, it should be noted that the transmission power consumption of the proposed algorithm is always **higher** than that of [Tao+16], this is due to the individual fronthaul capacity constraints are taken into account and respected here. Hence, the traffic load among each fronthauls are allocated according to their available resources, while in the algorithm proposed in [Tao+16], such regulations are ignored.

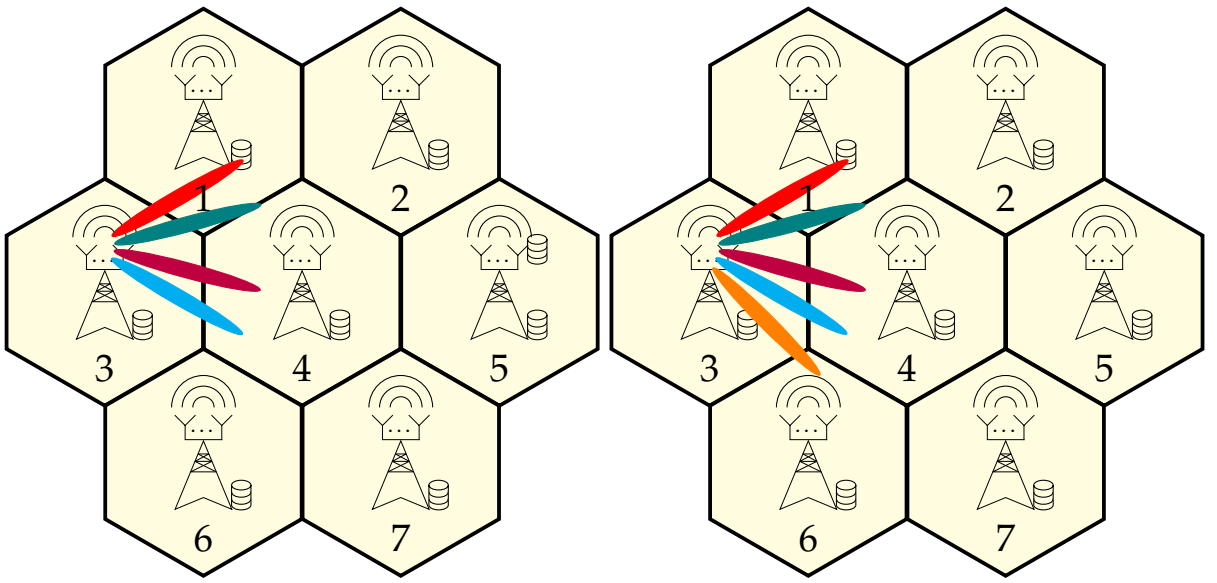
The results up to now only reflect the performance of a specific slot, the overall performance must also be investigated. In order to do this, 500 independent realizations are set up, i.e., 500 consecutive downlink slots are considered, and in each slot twelve UEs are randomly and independently selected within the network to be scheduled, each is with random content requests according to the Zipf distribution. The channel coefficients between UEs and eRRHs are also independently obtained using the channel model listed in Table 4.1. The proposed algorithm is then executed for optimizing the network for each downlink slot. The results of each realization are documented in terms of whether the network can be optimized to satisfy all UEs' demands, under the specific channel conditions of this slot, as well as the network resource configurations. In some realizations, it is infeasible to satisfy all UEs' requests. This is either due to many uncached contents happen to be requested, or the limited individual fronthaul capacities leading to non-sufficient cooperation between eRRHs so as to counteract the bad channel conditions within this slot. We compute the outage probability [6] based on 500 realizations and the results are depicted Fig. 4.16. It can be seen that with larger cache memory size

[5] As the algorithm proposed in [Tao+16] does not consider the individual fronthaul capacity constraints, we compute its average individual capacity for a fair comparison.

[6] It denotes the probability such that the QoS of each UE cannot be satisfied simultaneously under the channel conditions for the current slot.

(*a*) Cluster of eRRH 3 (Proposed)

(*b*) Cluster of eRRH 3 (Benchmark)



(*c*) Cluster of eRRH 5 (Proposed)

(*d*) Cluster of eRRH 5 (Benchmark)

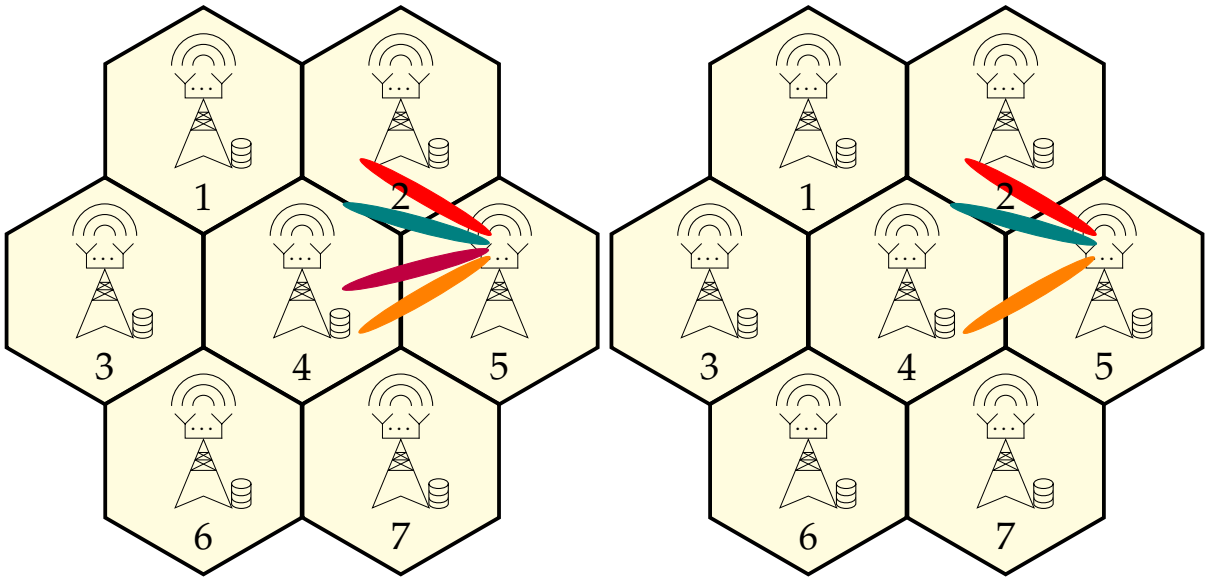


Figure 4.14: An illustration of the final cluster involvements of eRRH 3 and eRRH 5, which are obtained via the proposed Alg. 2 and the benchmark Alg. in [Tao+16]. Beams indicate that this eRRH is involved in the cluster to transmit the corresponding contents. Different beam colors denote different contents it shall transmit. The colors used here are in consistency with the legends used in Fig. 4.10 - Fig. 4.13, for distinguishing different contents.

Figure 4.15: The comparison of the network TX power consumption. *Benchmark scheme: Full cooperation between all eRRHs for all multi-cast groups. Case 1:* 70 *Mbps*, 2 *Contents Cached; Case 2:* 104 *Mbps*, 2 *Contents Cached, Case 3:* 104 *Mbps*, 3 *Contents Cached*.

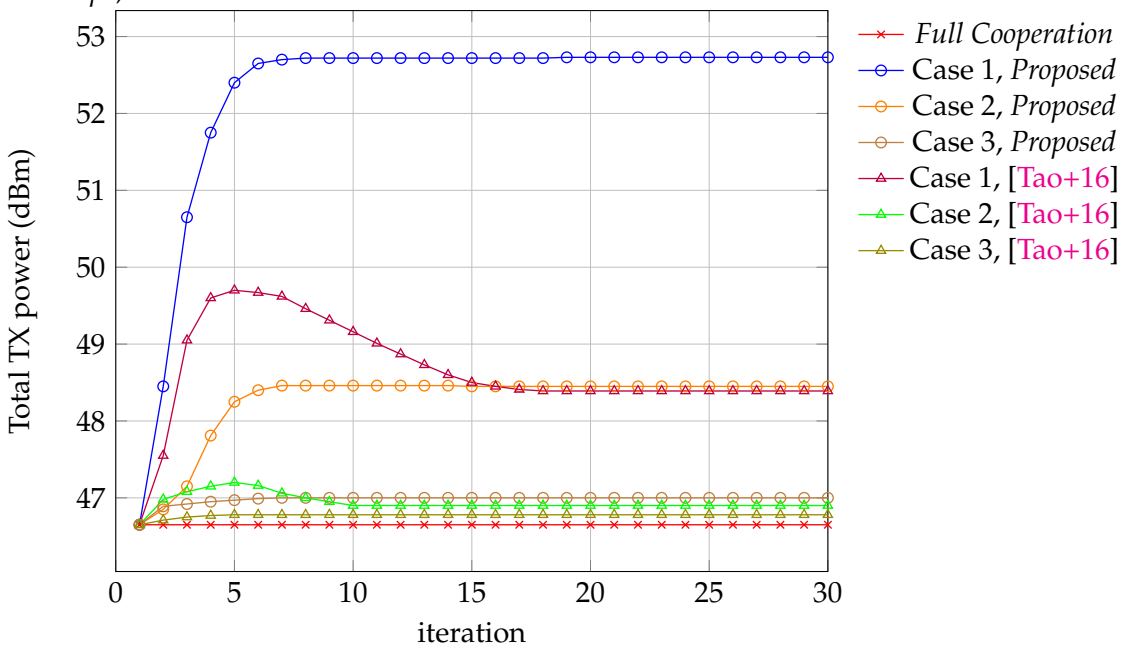


Figure 4.16: The outage probabilities for different fronthaul capacities and cache memory sizes.

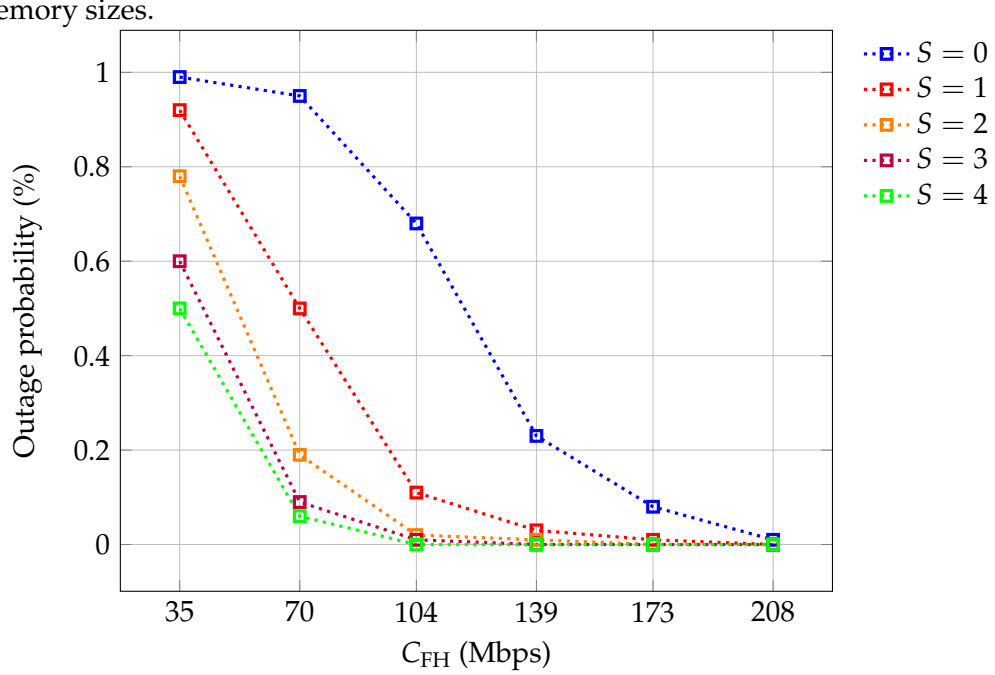
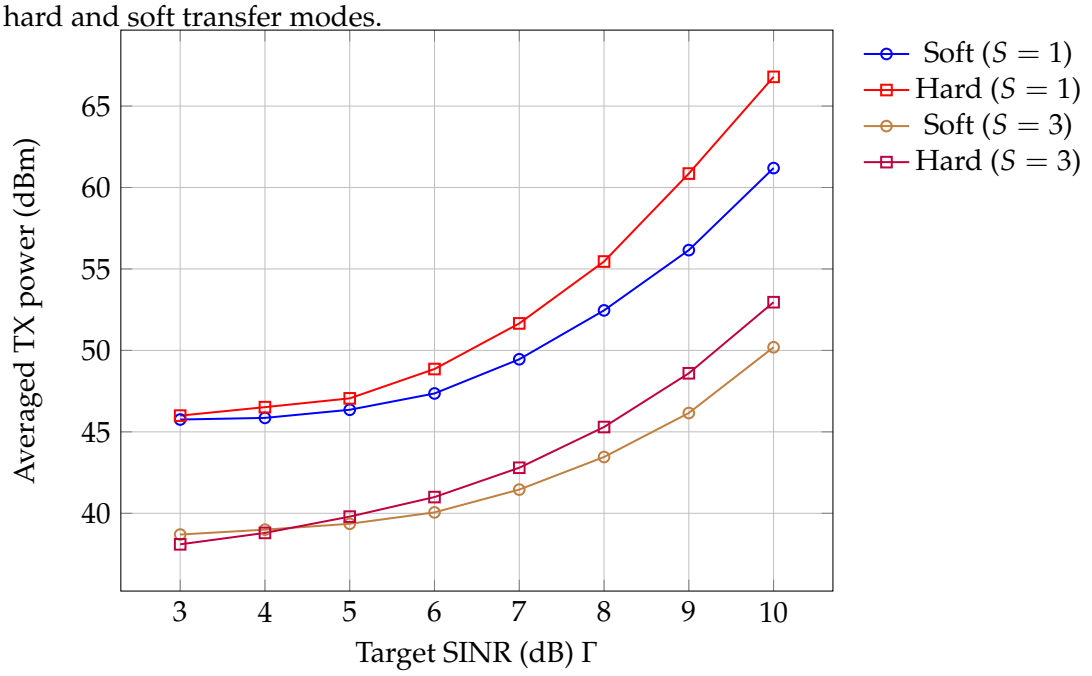


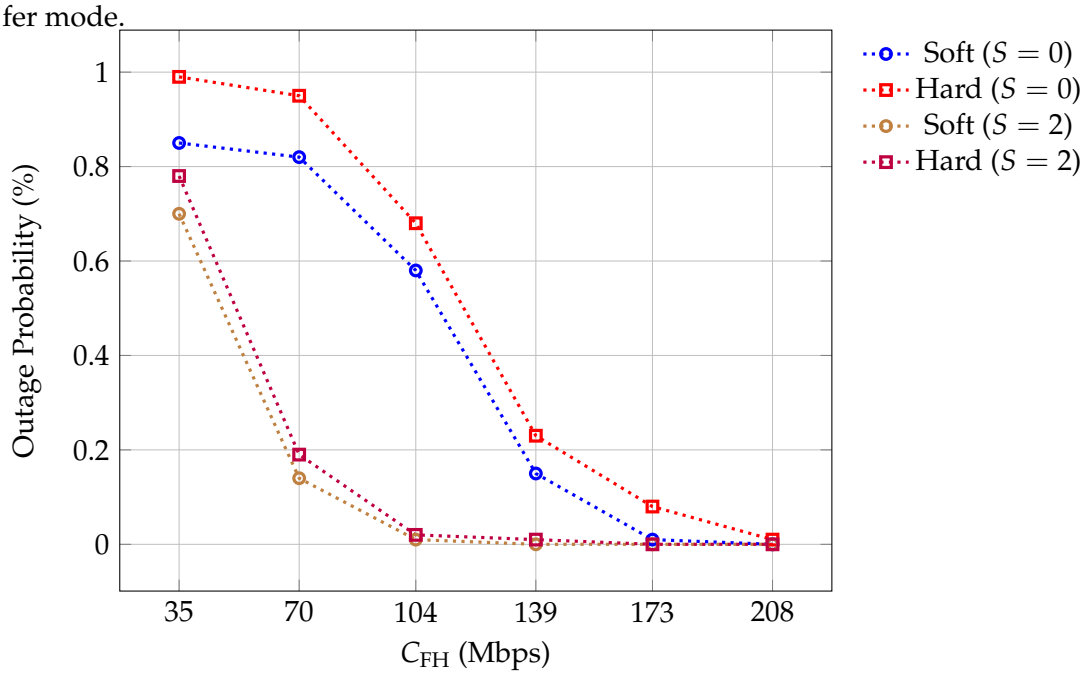
Figure 4.17: The minimized TX power obtained via the proposed algorithms for the hard and soft transfer modes.



and larger fronthaul capacity, the outage probability can be significantly reduced, because the transmission cooperation among more eRRHs becomes possible: If eRRHs have larger cache memory sizes, more contents can be cached without consuming the fronthaul resources, then more eRRHs can participate in transmitting these contents, which leads to higher spatial diversity either to decrease the transmission power, or to counteract the bad channel conditions. When the network has larger fronthaul capacity, the uncached contents can be delivered to more eRRHs, which also increase the possibility of the cooperation.

Next, the results for the soft transfer mode are collected. Similar to the simulation method introduced above, for each network configuration, i.e., any specific target SINR Γ and cache memory size S, we also set up 500 independent realizations: 500 consecutive downlink slots with independent and different scheduled UEs, channel conditions, requested contents, etc.. Then we adopt the proposed algorithms for both transfer modes and document the resultant minimized TX power of each realization. Finally, the obtained results are averaged and plotted in Fig. 4.17. The x-axis denotes different values of target SINRs and the y-axis denotes the minimized network TX power, which is averaged over 500 realizations. It can be observed that in most cases, the soft transfer mode is superior to the hard transfer mode, in terms of the TX power. Moreover, when the target SINR Γ becomes higher, or the cache memory size S becomes smaller, the gap between them becomes more prominent. The rationale of such a behaviour is easy to discover: Compared with the hard transfer mode, the soft transfer mode has higher data delivery efficiency from the BBU pool to each eRRH, as the compression is performed before sending them. In

Figure 4.18: The comparison of the outage probabilities for the hard and soft transfer mode.



contrast, for the hard transfer mode, raw data streams almost without any processing are sent to eRRHs, which has lower efficiency in terms of the utilization of the fronthaul resources. When the fronthaul resources can be exploited more efficiently, a specific uncached content can be delivered to more eRRHs, leading to higher spatial diversity and thus lower transmission power. Such an efficiency gap becomes more apparent, when the fronthaul resources becomes scarcer: For example, when the cache memory size gets smaller, less contents can be cached, thus more contents have to be fetched remotely via the fronthauls. Another example is when the target SINR gets larger, more eRRHs have to participate in each cluster to increase the spatial diversity to generate narrower beams for higher achievable SINRs. Hence, the uncached contents have to be delivered to more eRRHs. In both cases, more fronthaul resources are required, the advantage of the soft transfer mode over the hard one, in terms of exploiting the fronthaul resources, becomes more prominent. However, when the fronthaul resources are abundant, e.g., when $\Gamma = 3$ dB and $S = 3$, the TX power of the soft transfer mode is even higher. This is due to the quantization error introduced by the soft transfer mode, see (4.14). So it can be concluded that if the fronthaul resources are not the performance bottleneck, the introduced quantization error from the soft transfer mode might counteract its advantage. Such results can give some insights and be generalized to some guidelines when a real F-RAN is set up.

In Fig. 4.18, the outage probabilities are compared between these two transfer modes, for different fronthaul capacities and cache memory sizes. At first, it must be emphasized that such a comparison (actually also including the comparison in

Fig. 4.17) is a little unfair to the soft transfer mode: Remember that we have stated in the last paragraph of Subsection 4.1.5, the soft transfer mode has the potential to use less dedicated capacity for transmitting pilots, than the hard one, since only the precoders for the cached contents are to be transmitted via pilots. Hence, when these two schemes are compared with the same available fronthaul capacity (after deducting the dedicated capacity for pilots), the soft transfer mode actually requires less total fronthaul capacity than the hard transfer mode. However, the soft scheme still outperforms the hard one under such an *unfair* comparison. It can be seen from Fig. 4.18, due to higher data transmission efficiency, the soft transfer mode can exploit the available network resources better. Hence, it can achieve lower outage probability, or in other words, less probable to fail to serve all UEs with the target QoS, especially when the resources are limited. When the fronthaul capacity or the cache memory size gets larger, the gap between them becomes smaller.

Furthermore, it should be noted that the soft transfer mode has higher complexity, in both network operation and optimization: The BBU pool needs to multiplex and modulate the data streams and then performs the compression, and the eRRHs must do decompression in order to reconstruct the uncached data stream before being sent the to UEs. For the cached contents, eRRHs have to perform similar signal processing procedures compared with the hard transfer mode. Moreover, by comparing Alg. 2 and Alg. 3, we see that the proposed algorithm for the soft transfer mode has higher complexity, as more parameters are to be optimized, and more constraints exist.

By investigating Fig. 4.16 - Fig. 4.18, it is also worth to mention that, especially from the practical point of view, increasing the cache memory size, increasing the fronthaul capacity, or reducing the target QoS have similar effect on the reduction of the TX power or the outage probability. This is due to the fact that, all of them make more cooperation between eRRHs easier to happen. In practice however, the target QoS cannot be adjusted easily, and the deployment of the fronthaul with higher capacity is quite expensive and difficult. Hence, cache is a quite cheap and easy way to improve the overall performance, which can be a useful hint to the network providers.

4.2.2 High EE oriented Design — Total Power Minimization

In the previous subsection, only the transmission power of the network is minimized. Hence, all eRRHs must be active to achieve the highest spatial diversity for reducing the transmission power. However, as we have introduced in Subsection 4.1.3, the operational power of an eRRH, including the power consumed by circuits, cooling system and an active fronthaul, might be much higher than its transmission

power. If the total power consumption of the network is considered at the system level, activating all eRRHs to lower only the transmission power might not pay off, as much more operational power can be consumed. Hence, it is worth to investigate whether the network can be optimized in terms of not only the transmission power, but also the operational power. The results can tell the network providers: Is it possible to switch off some eRRHs to save power, especially at the off-peak time, while the remaining ones can still fulfill the service requirements. In this subsection, we are going to propose the corresponding algorithms to answer this question.

4.2.2.1 Problem Formulation and Solving Procedures

The problem formulation for minimizing the total power of the network is straightforward, as the constraints are same as (4.22)-(4.25) (for the hard transfer mode), or (4.47)-(4.50) (for the soft transfer mode). The difference lies only at the objective function: The operational power of an active eRRH should be taken into account. By adopting the power model described in Subsection 4.1.3, the problem for the hard transfer mode can be formulated as follows:

$$\mathcal{P}_{\text{Hard original}}: \min_{\{\mathbf{v}^m\}_{m=1}^M} \frac{1}{\xi}\left(\sum_{m=1}^{M}||\mathbf{v}^m||_2^2\right) + \sum_{n=1}^{N} P_{\text{o}}\left|\sum_{m=1}^{M}||\mathbf{v}_n^m||_2^2\right|_0 + \sum_{n=1}^{N} P_{\text{sleep}}\left(1-\left|\sum_{m=1}^{M}||\mathbf{v}_n^m||_2^2\right|_0\right), \tag{4.73}$$

$$\text{s.t.} \quad \text{SINR}_k^{\text{hard}} \geq \Gamma^m, \quad \forall k \in \mathcal{G}^m, \quad \forall \mathcal{G}^m, \tag{4.74}$$

$$\sum_{m=1}^{M}(1-c_n^{f^m})\left|||\mathbf{v}_n^m||_2^2\right|_0 \log_2(1+\Gamma^m) \leq C_{\text{FH},n} \quad \forall n \in \mathcal{N} \qquad \textit{dedicated}, \tag{4.75}$$

$$\sum_{n=1}^{N}\sum_{m=1}^{M}(1-c_n^{f^m})\left|||\mathbf{v}_n^m||_2^2\right|_0 \log_2(1+\Gamma^m) \leq C_{\text{FH}} \qquad \textit{non-dedicated}, \tag{4.76}$$

$$\sum_{m=1}^{M}||\mathbf{v}_n^m||_2^2 \leq P_{\text{TX},n}^{\max}, \quad \forall n \in \mathcal{N}. \tag{4.77}$$

For the soft transfer mode, it is as follows:

$$\mathcal{P}_{\text{Soft original}}: \min_{\{\mathbf{w}^m\}_{m=1}^M, \mathbf{q}} \frac{1}{\xi}\left(\sum_{m=1}^{M}||\mathbf{w}^m||_2^2 + \sum_{n=1}^{N}||\mathbf{q}_n||_2^2\right)$$
$$+\sum_{n=1}^{N} P_{\text{o}}\left|\sum_{m=1}^{M}||\mathbf{w}_n^m||_2^2\right|_0 + \sum_{n=1}^{N} P_{\text{sleep}}\left(1-\left|\sum_{m=1}^{M}||\mathbf{w}_n^m||_2^2\right|_0\right), \tag{4.78}$$

$$\text{s.t.} \quad \text{SINR}_k^{\text{soft}} \geq \Gamma^m, \quad \forall k \in \mathcal{G}^m, \quad \forall \mathcal{G}^m, \tag{4.79}$$

$$\sum_{l=1}^{L}\log_2\left(1+\frac{\sum_{m=1}^{M}(1-c_n^{f^m})|w_{n,l}^m|^2}{q_{n,l}^2}\right) \leq C_{\text{FH},n} \quad \forall n \in \mathcal{N} \qquad \textit{dedicated}, \tag{4.80}$$

$$\sum_{n=1}^{N}\sum_{l=1}^{L}\log_2\left(1+\frac{\sum_{m=1}^{M}(1-c_n^{f^m})|w_{n,l}^m|^2}{q_{n,l}^2}\right) \leq C_{\text{FH}} \qquad \textit{non-dedicated}, \tag{4.81}$$

$$\sum_{m=1}^{M}||\mathbf{w}_n^m||_2^2 + ||\mathbf{q}_n||_2^2 \leq P_{\text{TX},n}^{\max}, \quad \forall n \in \mathcal{N}. \tag{4.82}$$

The objective expression (4.73) denotes the total power consumption for the hard transfer mode. The first term of it indicates the total power consumption related to transmission, where $\sum_{m=1}^{M}||\mathbf{v}^m||_2^2$ denotes the total transmission power, and $\xi \in (0,1)$ denotes the power amplifier efficiency, please refer to Subsection 4.1.3 for more details. The second term denotes the total operational power consumption of all active eRRHs, and the third term indicates the total power consumption of all inactive eRRHs (if any). Remember that the ℓ_0-norm $\left|||\mathbf{v}_n^m||_2^2\right|_0$ has been adopted to denote whether eRRH n is involved into transmitting the content requested by multi-cast group $\mathcal{G}^m$. By executing the proposed algorithm, the optimized beamformer $\mathbf{v}_n^{m,\text{opt}}$ can be obtained, from which we finally know if eRRH n should serve multi-cast group $\mathcal{G}^m$ by computing the value of $\left|||\mathbf{v}_n^{m,\text{opt}}||_2^2\right|_0$ [7]. Here, the same technique can be utilized: By summing up all multi-cast groups that eRRH n serves, i.e., $\sum_{m=1}^{M}||\mathbf{v}_n^m||_2^2$, the total transmission power of this eRRH is derived. Hence, the ℓ_0-norm of it can be used to indicate whether it is active or not. When eRRH n should be involved in serving **at least** one multi-cast group, $\left|\sum_{m=1}^{M}||\mathbf{v}_n^m||_2^2\right|_0$ is 1, meaning that it must be activated and the operational power P_{o} is consumed. Otherwise, the value of the ℓ_0-norm is 0, this eRRH can be deactivated and only power P_{sleep} in sleep mode is consumed. If this problem can be solved, the value of $\left|\sum_{m=1}^{M}||\mathbf{v}_n^{m,\text{opt}}||_2^2\right|_0$ can tell (together with the unit step function as we have introduced before) whether eRRH n can be switched off to save more power.

Similarly, the objective expression (4.78) for the soft transfer mode adopts the same method. The only difference is that the precoders $\{\mathbf{w}^m\}_{m=1}^M$ are designed at the BBU pool, and a part of the transmission power $\sum_{n=1}^{N}||\mathbf{q}_n||_2^2$ is consumed by the quantization noise introduced by the compression.

[7] As introduced in Subsection 4.2.1.1, we use the unit step function to determine whether $\left|||\mathbf{v}_n^{m,\text{opt}}||_2^2\right|_0$ is 0 or 1 with the predetermined threshold parameter τ.

Both objectives make deactivating some eRRHs possible: If the operational power saved by deactivating an eRRH can compensate the increased transmission power among all others (as the aggregated array gain/spatial diversity is decreased), and the remaining eRRHs can still satisfy the QoS of each UE and fulfill other constraints, this eRRH shall be switched off. Namely, for both transfer modes, the decrease of the second terms of the objectives must lead to an increase of the first terms, and vice versa.

As the same methods have been adopted in formulating the problem of minimizing the total transmission power, and the constraints remain unchanged, the same techniques, i.e., SDR, the iterative ℓ_0-norm approximation, EVD, etc., can be utilized to solve the new problem. In order to avoid repetitions, here we only briefly introduce the solving procedures for the hard transfer with dedicated fronthaul, i.e., (4.73)-(4.75) and (4.77). Extensions to other cases is straightforward by referring to previous sections with minor modifications.

Note that the objective (4.73) can be equivalently written as

$$\min_{\{\mathbf{v}^m\}_{m=1}^M} \sum_{m=1}^{M} ||\mathbf{v}^m||_2^2 + \sum_{n=1}^{N} \Delta P \Big| \sum_{m=1}^{M} ||\mathbf{v}_n^m||_2^2 \Big|_0 + \xi N P_{\text{sleep}}, \tag{4.83}$$

where $\Delta P = \xi(P_{\text{o}} - P_{\text{sleep}})$. As the last term is a constant, it is sufficient to consider only the first and second terms as the equivalent objective. The first step is still to reformulate the problem to the form that SDR can be applied:

$$\mathcal{P}_{\text{Hard}}: \min_{\{\mathbf{V}^m\}_{m=1}^M} \quad \sum_{m=1}^{M} \text{tr}(\mathbf{V}^m) + \sum_{n=1}^{N} \Delta P \Big| \sum_{m=1}^{M} \text{tr}(\mathbf{V}^m \mathbf{J}_n) \Big|_0, \tag{4.84}$$

$$\text{s.t.} \quad \Gamma^m \left(\sigma_k^2 + \sum_{i \neq m}^{M} \text{tr}(\mathbf{V}^i \mathbf{H}_k) \right) - \text{tr}(\mathbf{V}^m \mathbf{H}_k) \leq 0, \quad \forall k \in \mathcal{G}^m, \quad \forall \mathcal{G}^m, \tag{4.85}$$

$$\sum_{m=1}^{M} (1 - c_n^{f^m}) \big| \text{tr}(\mathbf{V}^m \mathbf{J}_n) \big|_0 \log_2 (1 + \Gamma^m) \leq C_{\text{FH},n}, \quad \forall n \in \mathcal{N}, \tag{4.86}$$

$$\sum_{m=1}^{M} \text{tr}(\mathbf{V}^m \mathbf{J}_n) \leq P_{\text{TX},n}^{\max}, \quad \forall n \in \mathcal{N}, \tag{4.87}$$

$$\mathbf{V}^m \succeq \mathbf{0}, \quad \forall m \in \mathcal{M}, \tag{4.88}$$

$$\text{rank}\,(\mathbf{V}^m) = 1, \quad \forall m \in \mathcal{M}. \tag{4.89}$$

Obviously, the approximation of the ℓ_0-norm in the second term of objective (4.84) is required. This can also be achieved in an iterative manner, which is similar to what we have done with (4.33) and (4.34): In the $(t+1)$-th iteration, $\big| \sum_{m=1}^{M} \text{tr}(\mathbf{V}^{m(t+1)} \mathbf{J}_n) \big|_0$ is approximated as a linear function of $\sum_{m=1}^{M} \text{tr}(\mathbf{V}^{m(t+1)} \mathbf{J}_n)$ as

$$\big| \sum_{m=1}^{M} \text{tr}(\mathbf{V}^{m(t+1)} \mathbf{J}_n) \big|_0 \approx u_n^{(t+1)} \sum_{m=1}^{M} \text{tr}(\mathbf{V}^{m(t+1)} \mathbf{J}_n), \tag{4.90}$$

where the re-weighted coefficient $u_n^{(t+1)}$ is calculated via the result of the previous iteration:

$$u_n^{(t+1)} = \frac{1}{\tau + \sum_{m=1}^{M} \mathrm{tr}(\mathbf{V}^{m(t)} \mathbf{J}_n)}. \tag{4.91}$$

With SDR, and dropping the constraint (4.89), the problem to be solved in $(t+1)$-th iteration can be formulated as

$$\mathcal{P}_{\text{Hard}}^{(t+1)} : \min_{\{\mathbf{V}^{m(t+1)}\}_{m=1}^{M}} \quad \sum_{m=1}^{M} \mathrm{tr}(\mathbf{V}^{m(t+1)}) + \Delta P \sum_{n=1}^{N} u_n^{(t+1)} \sum_{m=1}^{M} \mathrm{tr}(\mathbf{V}^{m(t+1)} \mathbf{J}_n), \tag{4.92}$$

$$\text{s.t.} \quad \Gamma^m \sum_{i \neq m}^{M} \mathrm{tr}(\mathbf{V}^{i(t+1)} \mathbf{H}_k) - \mathrm{tr}(\mathbf{V}^{m(t+1)} \mathbf{H}_k) + \Gamma^m \sigma_k^2 \leq 0, \quad \forall k \in \mathcal{G}^m, \quad \forall \mathcal{G}^m, \tag{4.93}$$

$$\sum_{m=1}^{M} a_n^{m(t+1)} \mathrm{tr}(\mathbf{V}^{m(t+1)} \mathbf{J}_n) - C_{\text{FH},n} \leq 0, \quad \forall n \in \mathcal{N}, \tag{4.94}$$

$$\sum_{m=1}^{M} \mathrm{tr}(\mathbf{V}^{m(t+1)} \mathbf{J}_n) \leq P_{\text{TX},n}^{\max}, \quad \forall n \in \mathcal{N}, \tag{4.95}$$

$$\mathbf{V}^{m(t+1)} \succeq \mathbf{0}, \quad \forall m \in \mathcal{M}, \tag{4.96}$$

where

$$a_n^{m(t+1)} = k_n^{m(t+1)} (1 - c_n^{f^m}) \log_2 (1 + \Gamma^m) \tag{4.97}$$

with $k_n^{m(t+1)}$ calculated according to (4.34).

The relaxed and reformulated problem above consists of a linear objective function, $K + 2N$ linear inequality constraints, and M positive-semidefinite constraints, which is also a standard SDP problem.

Similarly, an initial problem shall be formulated to obtain the initial values of $\{u_n\}_{n=1}^{N}$. We assume all eRRHs are activated in the very beginning, hence, the second term of (4.92) is temporarily dropped since minimizing the total power is equivalent to minimizing only the transmission power in this case. Therefore, the initial problem $\mathcal{P}_{\text{Hard}}^{(0)}$ is the same as (4.40)-(4.43), and the solving procedure is summarized in Alg. 4.

After the initial step, where all eRRHs are activated, the second term of (4.92) and the fronthaul constraint (4.94) are added again to formulate the problem for next iterations. Two re-weighted coefficient sets, i.e., $u_n^{(t+1)}$ and $k_n^{m(t+1)}$, $\forall m, n$, are amended gradually in each iteration. eRRH i might be switched off (deactivation) gradually, as long as its transmission power $P_{\text{TX},i} = \sum_{m=1}^{M} \mathrm{tr}(\mathbf{V}^m \mathbf{J}_i)$ falls below the threshold parameter τ. Similarly, an active eRRH j might be gradually excluded from cluster $\mathcal{C}^m$ for serving multi-cast group $\mathcal{G}^m$, when its corresponding power for

this multi-cast group $P_{\mathrm{TX},j,f(m)} = \mathrm{tr}(\mathbf{V}^m \mathbf{J}_j)$ falls below τ [8]. The objective of the minimization problem (4.92) and constraints (4.93)-(4.96) ensure that such deactivation and exclusion happen, only when the resultant total power consumption can be decreased, and the new clustering pattern can meet the QoS of each UE, the load on each fronthaul does not exceed its capacity, and the individual power constraint of each eRRH can be respected.

Algorithm 4: The Iterative Optimization Steps for Total Power Minimization (*For the hard transfer mode*)

Initialization: Solve the standard SDP problem $\mathcal{P}^{(0)}_{\mathrm{Hard}}$ (4.40)-(4.43) to obtain $\{\mathbf{V}^{m(0)}\}_{m=1}^{M}$. Compute $a_n^{m(1)}$ based on (4.97), $\forall m, n$, and the values of $u_n^{(1)}$ based on (4.91), $\forall n$. Construct the problem $\mathcal{P}^{(1)}_{\mathrm{Hard}}$ according to (4.92)-(4.96), and set $t \leftarrow 1$.
repeat
  Solve the standard SDP problem $\mathcal{P}^{(t)}_{\mathrm{Hard}}$ for obtaining $\{\mathbf{V}^{m(t)}\}_{m=1}^{M}$.
  Update the values of $a_n^{m(t+1)}$ based on (4.97), $\forall m, n$, and the values of $u_n^{(t+1)}$ based on (4.91), $\forall n$. Then formulate the problem $\mathcal{P}^{(t+1)}_{\mathrm{Hard}}$ according to (4.92)-(4.96), and set $t \leftarrow t + 1$.
until *convergence or reaching the max iteration number*;
if $\mathrm{rank}(\mathbf{V}^{m(\mathrm{last})}) = 1$ **then**
  Perform EVD to obtain the optimal $\{\mathbf{v}^m\}_{m=1}^{M}$.
else
  Use Gaussian randomization and scaling [KSL08] method to obtain the approximate solution $\{\mathbf{v}^m\}_{m=1}^{M}$.

Extension to the soft transfer mode and non-dedicated fronthaul: Such extensions are straightforward. Extension to the scenario of the non-dedicated fronthaul is the same as what introduced in Subsection 4.2.1.1. For the extension to the soft transfer mode, we only need to combine the technique introduced in Subsection 4.1.4.2, with the ℓ_0-norm iterative approximation method introduced above to convexify the second and third term of (4.78). Alg. 4 can be amended in a straightforward way to solve the resultant problem.

4.2.2.2 Numerical Results

In this subsection the numerical results of the proposed algorithms are to be provided via the simulations. The network setup and the simulation environment are the same as the description in Subsection 4.2.1.3. The same simulation parameters listed in Table 4.1 are adopted, but the total power consumption, instead of only

[8] We can set different values of the threshold parameter τ used for these two iterative approximation procedures. Here we select the same value for simplicity.

transmission power as previously, will be documented. Moreover, the power amplifier efficiency ξ of each four-antenna eRRH is set to be 0.25, and each active eRRH is assumed to consume 35 Watt to maintain its operation, i.e., $P_{\text{o}} = 35$ W. The eRRH in sleep mode is suppose to consume 5 Watt for monitoring potential commands, i.e., $P_{\text{sleep}} = 5$ W. The results are to be compared with the ones proposed in [Tao+16], which is set as a benchmark. In the benchmark algorithm, only the transmission power is minimized, and individual fronthaul capacity constraints, as well as the operational power of an active eRRH, are not considered.

Two representative scenarios are selected to illustrate the results respectively.

Representative Scenario 1 (Abundant local resources): In this specific downlink slot, after twelve scheduled UEs submit their requests according to the Zipf distribution (4.1), the BBU pool finds that only four different contents are requested, and three of them have been already cached at eRRHs. In scenarios like this, i.e., most requested contents have already been available at local caches without the need of being delivered remotely from the cloud via fronthauls, there are comparatively sufficient caching and fronthaul capacity resources. For the proposed and benchmark algorithms, the transmission power of eRRH n, $P_{\text{TX,n}} = \sum_{m=1}^{M} ||\mathbf{v}_n^m||_2^2, \forall n \in \mathcal{N}$, are recorded and illustrated in Fig. 4.19 and Fig. 4.20. The total power consumed by transmission, which is $\frac{1}{\xi} P_{\text{TX,tot}} = \frac{1}{\xi} \sum_{n=1}^{N} P_{\text{TX,n}}$, and the total operational power consumption, $P_{\text{o,tot}} = \sum_{\text{active eRRHs}} P_{\text{o}} + \sum_{\text{inactive eRRHs}} P_{\text{sleep}}$ are shown in Fig. 4.22. The total power consumption by computing $P_{\text{tot}} = \frac{1}{\xi} P_{\text{TX,tot}} + P_{\text{o,tot}}$ is shown in Fig. 4.23.

Representative Scenario 2 (Limited local resources): In another downlink slot, after twelve scheduled UEs submit their requests, unfortunately, seven different contents are requested, only three of them are cached. In scenarios like this, i.e., most requested contents have to be delivered via fronthauls, there are comparatively tight and limited caching and fronthaul capacity resources. Thus, less cooperative transmission is expected. Similar power comparisons are shown in Fig. 4.24 - Fig. 4.28.

Analysis of Fig. 4.19 - Fig. 4.23: We firstly discuss Representative Scenario 1 (abundant local resources), and the corresponding results acquired via the proposed algorithm. As most requested contents have been cached, abundant caching resources result in low traffic load on fronthauls. Thus, there are sufficient fronthaul capacities for the delivery of the uncached contents to as many eRRHs as possible. In another word, it is easy to form larger clusters to achieve more cooperation between eRRHs so as to decrease the transmission power. Therefore, switching off some eRRHs and fronthauls for saving the operational power is more probable, as the remaining eRRHs can still fulfill the UEs' demands. As we see from Fig. 4.19, the transmission power of three eRRHs, i.e., eRRH 1, eRRH 5 and eRRH 7, fall below -150 dBm in less than ten iterations, meaning that these eRRHs are determined to be switched

Figure 4.19: Representative Scenario 1: eRRH Deactivation (*Proposed*)

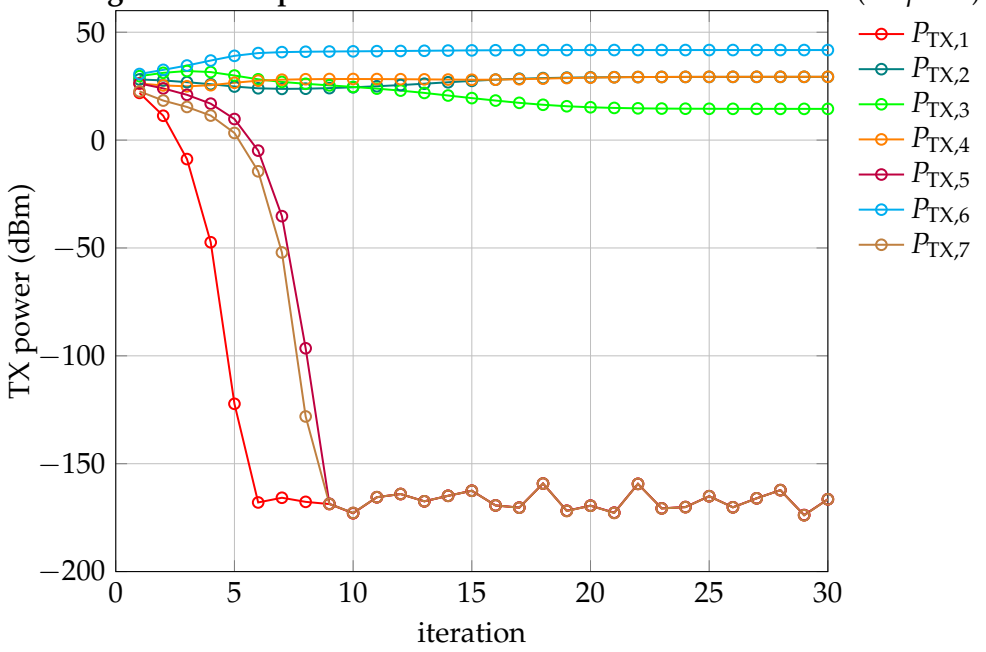
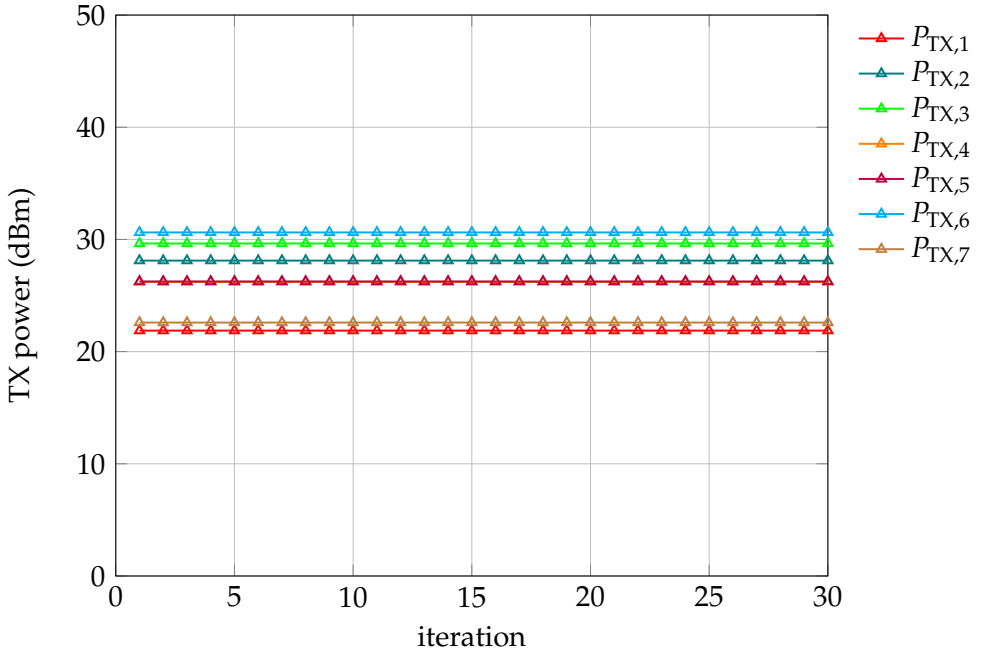


Figure 4.20: Representative Scenario 1: eRRH Deactivation (*Benchmark*)



(*a*) eRRH Deactivation (Proposed) (*b*) eRRH Deactivation (Benchmark)

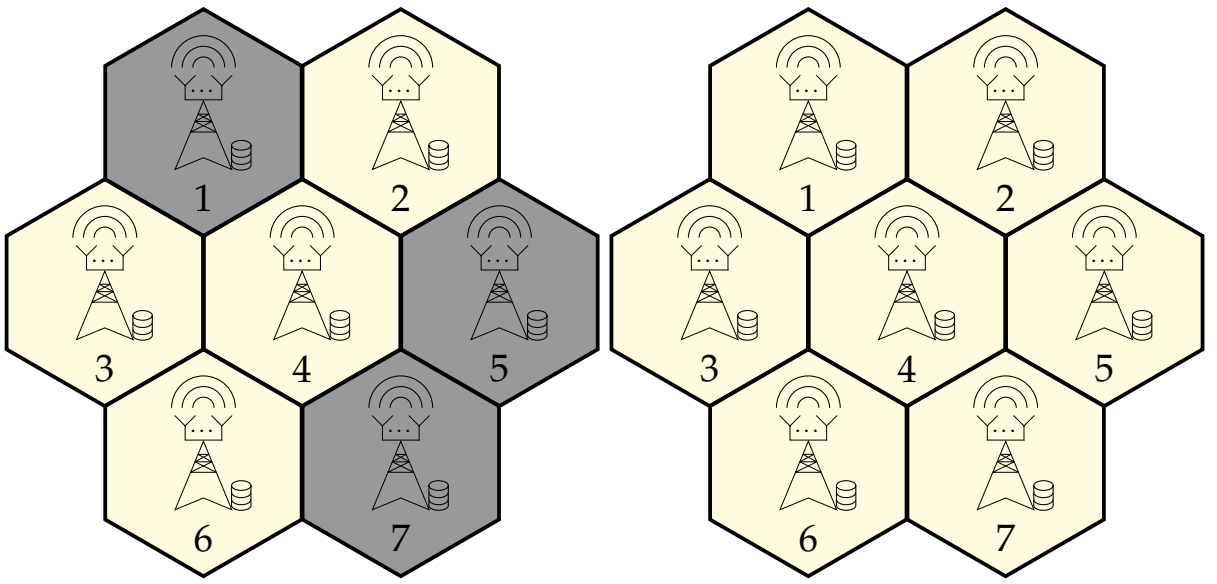


Figure 4.21: An illustration of the final eRRH deactivation results of **Representative Scenario 1**, with the proposed Alg. 4 and the benchmark Alg. in [Tao+16]. Cell colored with gray denotes that the eRRH within this cell is deactivated.

Figure 4.22: Representative Scenario 1: Power Evolution

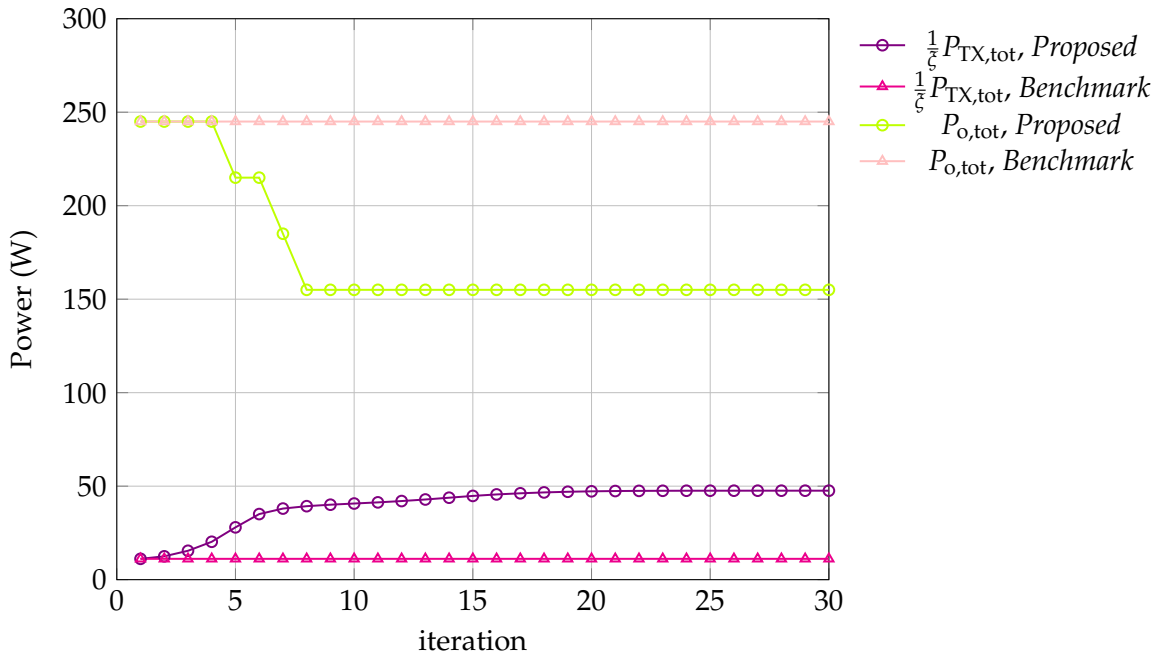
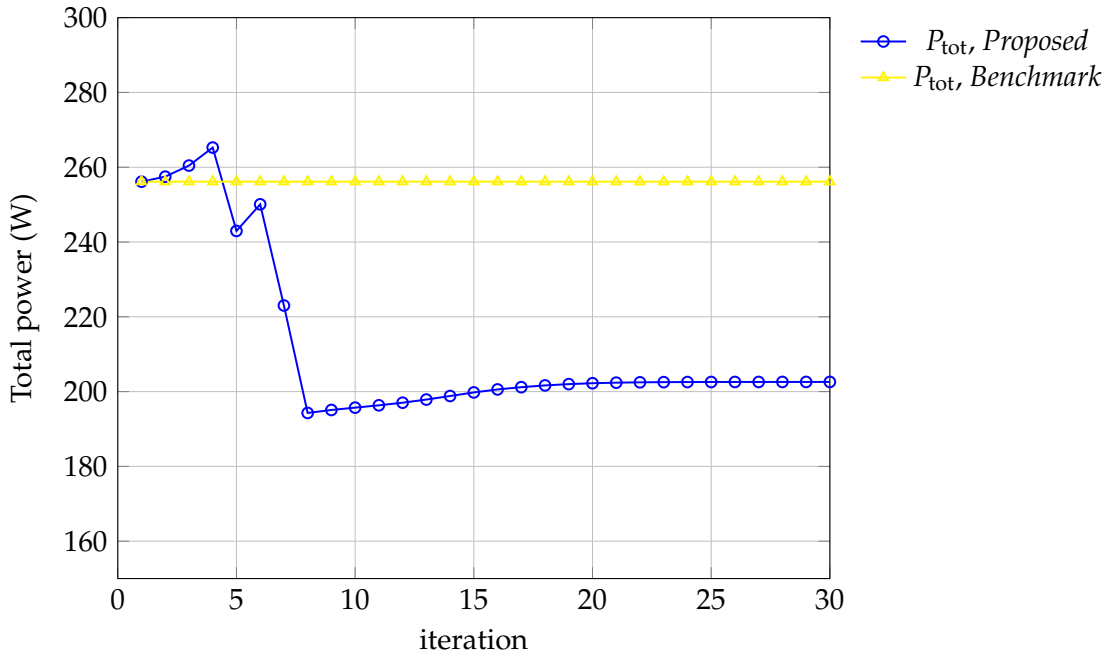


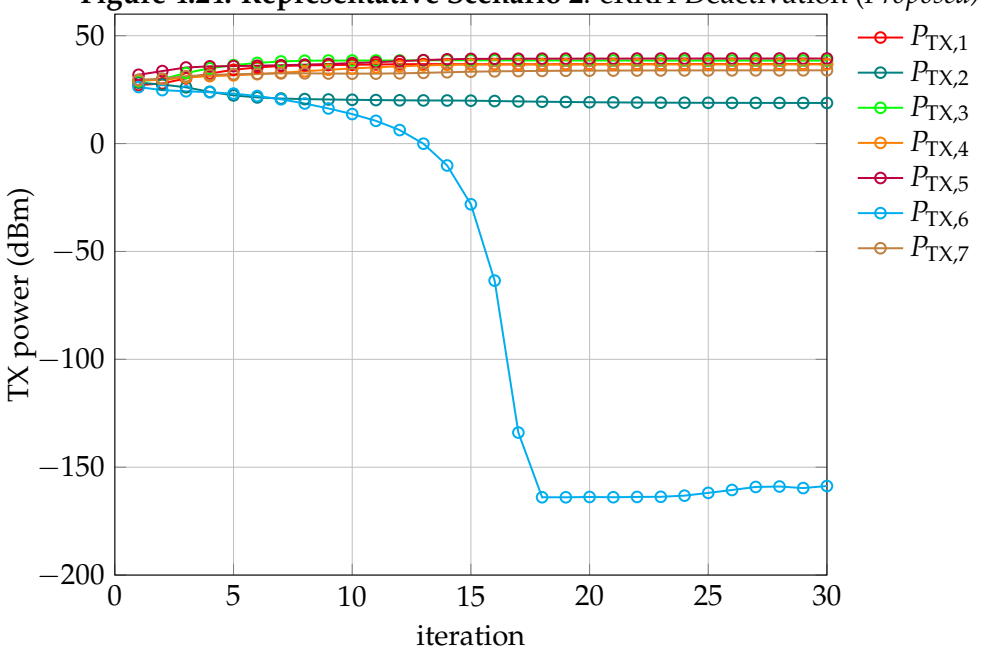
Figure 4.23: Representative Scenario 1: Total Power Consumption



off by the BBU pool via executing our proposed algorithm. For better illustration, the eRRH deactivation results of the proposed algorithm are depicted in Fig. 4.21 (a). Due to the deactivation behaviour, in Fig. 4.22, the total operation power $P_{\text{o,tot}}$ drops. The total power consumed by transmission, i.e., $\frac{1}{\xi}P_{\text{TX,tot}}$, continuously increases in the first seven iterations, then it converges into a relatively steady state. The reason behind is straightforward: Remember that we start with solving the initial problem by temporarily dropping individual capacity constraints (4.94), and the term of the operational power in the objective expression (4.92), in order to obtain the initial values for ℓ_0-norm approximation. In next iterations, they are added, and the re-weighted coefficients are computed and amended in each iteration. Several eRRHs are gradually forced to be switched off, or be excluded from some specific clusters. Hence, the transmission power from the proposed algorithm in Fig. 4.22 is increased mainly due to the two factors described above, i.e., 1. The individual fronthaul capacity constraints are added; 2. Less potential aggregated array gain resulting from deactivation of some eRRHs. Moreover, after about ten iterations, the proposed algorithm converges and reaches a stable phase.

Now we discuss the results obtained by the benchmark scheme. As only the transmission power is minimized, all eRRHs are kept to be active for increasing the potential spatial diversity to reduce the transmission power, as shown in Fig. 4.20. As a comparison to the proposed algorithm, the eRRH deactivation results of the benchmark algorithm are also plotted in Fig. 4.21 (b). Since no eRRH is deactivated, the total operational power remains the same, as shown in Fig. 4.22. This is also true for the total power consumption shown in Fig. 4.23, the results from the benchmark do not vary significantly, as both the operational power and the trans-

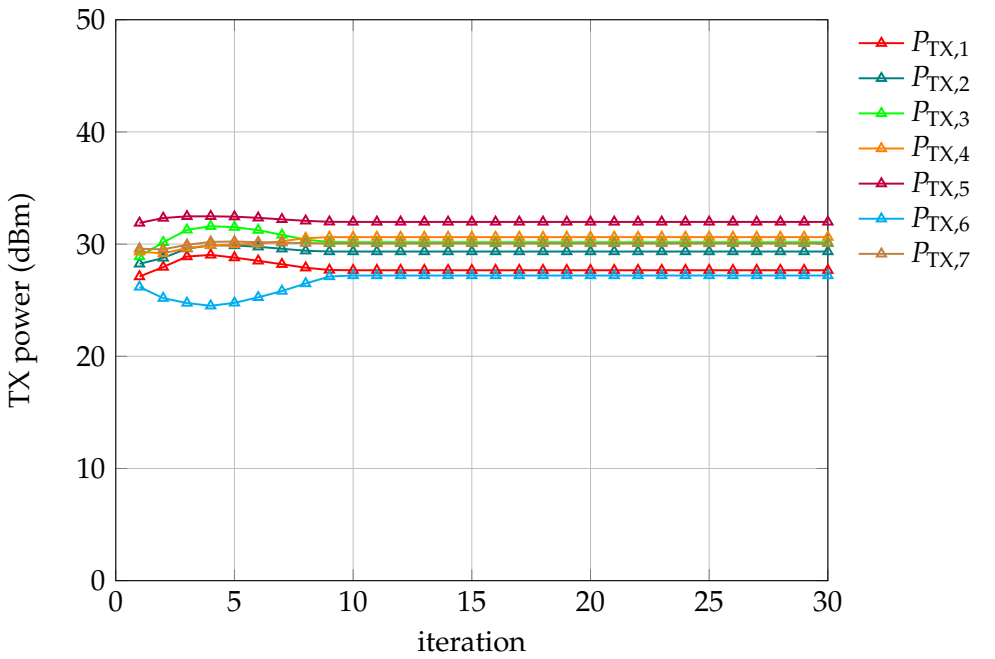
Figure 4.24: Representative Scenario 2: eRRH Deactivation (*Proposed*)



mission power among all eRRHs stay nearly unchanged. From this figure, we can observe the effectiveness of the proposed algorithm in saving the network power when the operational power is considered. However, we must say that the total power consumption with the proposed algorithm is not always less than that of the benchmark, as a more stringent problem to balance the traffic load on each active fronthaul is solved by us. Hence, in scenarios where less or even no eRRHs can be deactivated (e.g., Representative Scenario 2), the total power consumption might be higher than that of the benchmark algorithm, as we show next.

Analysis of Fig. 4.24 - Fig. 4.28: The results of Representative Scenario 2 are depicted in these figures, where most requested contents have to be delivered via fronthauls, resulting in heavy traffic load on them. The fronthaul capacity becomes a bottleneck, and forming larger clusters for more cooperation between eRRHs becomes more difficult. In such scenarios, with the proposed algorithm, it can be observed that only eRRH 6 and its fronthaul can be switched off after 16 iterations, as shown in Fig. 4.24. In Fig. 4.25, all eRRHs are still active as the operational power consumption is not considered in the benchmark scheme. Such deactivation behaviours are also plotted in Fig. 4.26 in a more intuitive way. Furthermore, Fig. 4.27 shows that the transmission power of the both algorithms increase after adding capacity constraints from the second iteration. However, the increasing rate of the proposed algorithm is much higher, due to the necessity of balancing the traffic load on very limited fronthaul resources. Hence, although one eRRH is switched off with the proposed algorithm, the total power consumption of it is still higher than the benchmark based on this unfair comparison, as shown in Fig. 4.28. Despite higher total power consumption, the traffic on each fronthaul is guaranteed to be supported. It

Figure 4.25: Representative Scenario 2: eRRH Deactivation (*Benchmark*)



(*a*) eRRH Deactivation (Proposed)

(*b*) eRRH Deactivation (Benchmark)

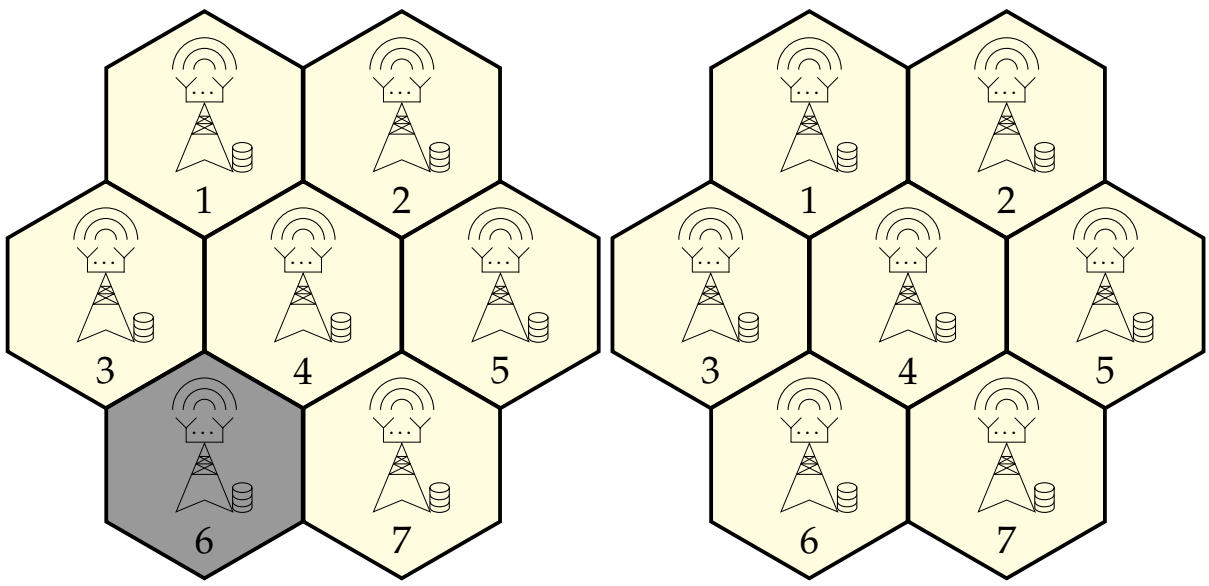


Figure 4.26: An illustration of the final eRRH deactivation results of **Representative Scenario 2**, with the proposed Alg. 4 and the benchmark Alg. in [Tao+16]. Cell colored with gray denotes that the eRRH within this cell is deactivated.

Figure 4.27: Representative Scenario 2: Power Evolution

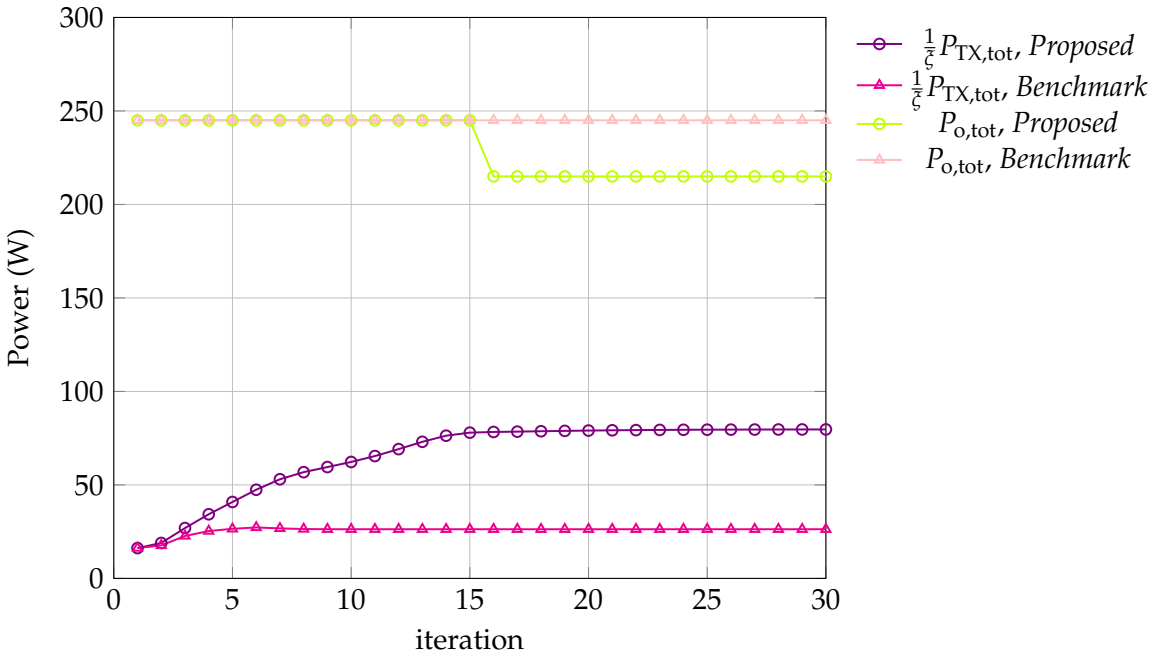


Figure 4.28: Representative Scenario 2: Total Power Consumption

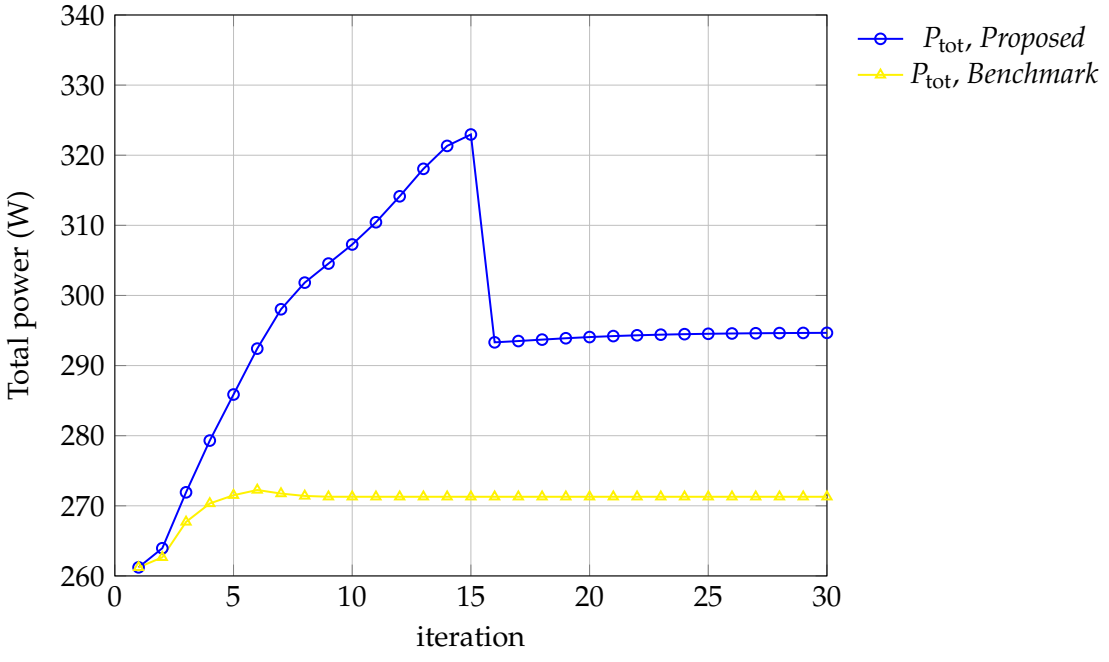
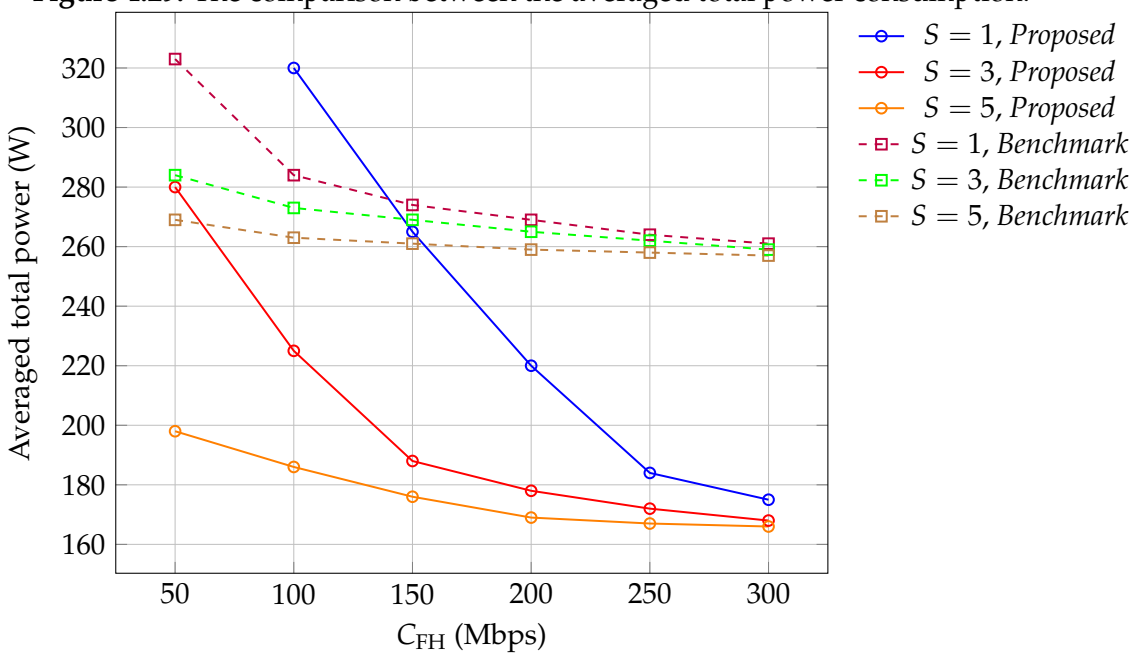


Figure 4.29: The comparison between the averaged total power consumption.



is also worth to mention that in Fig. 4.28, the sharp drop of the total power with the proposed algorithm is due to the deactivation of eRRH 6 around iteration 16, as also shown in Fig. 4.24.

Analysis of Fig. 4.29: At last, we configure the network such that eRRHs have variable individual fronthaul capacities and cache memory sizes. For each configuration, 300 independent realizations are set up, and the resultant total power consumption of the proposed and benchmark algorithm are documented. By averaging the results for each specific network configuration, Fig. 4.29 is acquired. It can be seen that by increasing either the individual fronthaul capacity or the cache memory size, the power consumption resulting from both algorithms decrease. Larger C_{FH} or S can make the local network resources more abundant, thus more cooperation becomes possible, and the realizations similar to Representative Scenario 1 is also more probable: More eRRHs are possible to be switched off with the proposed algorithm, leading to far less total power consumption than the benchmark scheme. However, when C_{FH} or S is smaller, it is likely that more realizations works in scenarios similar to Representative Scenario 2: Less or even no eRRHs can be switched off, and the traffic handling on active fronthauls becomes an significant issue. The load balancing makes the solution of proposed algorithm consume more total power than that of the relaxed benchmark problem. Furthermore, when C_{FH} and S get large enough, the solution of both algorithms enter the *saturation region*, since the current network resources have been sufficient to allow full cooperation for most requests, increasing local resources further cannot further increase the possibility of cooperation in order to further decrease the power significantly. With the proposed algorithm, it is also not possible to deactivate more eRRHs. However, it

shows that the power consumption in this region can be greatly reduced with the proposed algorithm, compared with the benchmark, due to the huge operational power saved by deactivation. Moreover, we emphasize again that increasing the cache memory is usually much easier and cheaper compared with increasing the fronthaul capacity.

4.2.3 High SE oriented Design — wMMF Metric

After the intensive discussion of the high EE oriented design of the cache-enabled F-RAN, it is time to investigate the high SE oriented design. In this subsection and the next one, we are going to address the optimization strategies for high SE based on two distinct metrics: One concerns the multi-cast Throughput Maximization (**TP-Max**) of the network, and the other one concerns the (weighted) Max-Min Fairness (**wMMF**). The metric of the multi-cast throughput is easy to be understood, as high throughput is almost the synonym of high SE. However, maximizing the network throughput might render individual achievable rates far more different among UEs, especially when some UEs have poor channel qualities (e.g., at cell edges), as more network resources tend to be prioritized on UEs with good channel qualities, so as to fully exploit the limited resources for maximizing the throughput. In this case, the QoS of some UEs cannot be guaranteed. In order to avoid such unfairness, the metric of Max-Min Fairness (MMF) intends to maximize the minimized achievable rate. Moreover, it is possible to add weights to different UEs, addressing different significance and priorities. Obviously, although the network throughput of the second design target is not maximized, it guarantees a predetermined fairness among all scheduled UEs.

As we are going to see, the solving procedure of maximizing the (weighted) minimized achievable rate, can be derived from the solving procedure of the high EE oriented design introduced in the previous subsections. Therefore, we start with addressing the problem of wMMF. The solving procedure for maximizing the multi-cast throughput will be discussed in the next subsection, as it is more complicated and some new techniques are to be introduced. To avoid repetitions, we consider only **the hard transfer mode with dedicated fronthaul**, as the extensions to other cases are similar to the methods we have introduced before.

When wMMF is considered, the QoS of each UE is guaranteed to achieve some extent of fairness, according to its predetermined weighting coefficient. In our F-RAN model, the UE with the worst channel conditions within a multi-cast group determines achievable rate of the content requested by all UEs of this group. If a specific content need to be prioritized, the weighting coefficient of it has to be

selected carefully. The problem of weighted Max-Min rate fairness is formulated as follows:

$$\mathcal{F}_{\text{wMMF}}^{\text{original}}(\mathbf{s},\mathbf{P}): \quad \max_{\{\mathbf{v}^m\}_{m=1}^M} \min_{m\in\mathcal{M}} \min_{k\in\mathcal{G}^m} \frac{1}{s^m}\text{SINR}_k, \tag{4.98}$$

$$\text{s.t.} \quad \sum_{m=1}^{M} ||\mathbf{v}_n^m||_2^2 \leq P_n, \qquad \forall n \in \mathcal{N}, \tag{4.99}$$

$$\sum_{m=1}^{M} (1-c_n^{f^m}) \left|||\mathbf{v}_n^m||_2^2\right|_0 \log_2\left(1+\min_{k\in\mathcal{G}^m}\text{SINR}_k\right) \leq C_{\text{FH},n}, \qquad \forall n \in \mathcal{N}. \tag{4.100}$$

As the hard transfer mode is assumed to be adopted, the achievable SINR at UE k can be referred to (4.15). The same notations are used as before: $\mathbf{v}^m = [\{\mathbf{v}_1^m\}^H, \{\mathbf{v}_2^m\}^H, ..., \{\mathbf{v}_N^m\}^H]^H \in \mathbb{C}^{NL\times 1}$ denotes the aggregate beamforming vector for content f^m requested by multi-cast group $\mathcal{G}^m$, and $\mathbf{v}_n^m$ denotes the part of this beamformer that is constructed at eRRH n. The scaling vector $\mathbf{s} = [s^1, s^2, ..., s^M]^T$ consists of M predetermined weighting coefficients for different contents that are requested by M multi-cast groups. In the objective function (4.98), we see two min operations: One is for UEs within the multi-cast group $\mathcal{G}^m$, and the other is among all existing multi-cast groups. The first one, i.e., $\min_{k\in\mathcal{G}^m}$, is due to the worst UE determines the achievable rate of the content requested by its multi-cast group. The second one, i.e., $\min_{m\in\mathcal{M}}$, together with the weighting coefficients s^m, aims to achieve the predetermined weighted fairness among all requested contents. For the content requiring higher QoS expectation at the UE side, its weighting coefficient is set to be larger. Hence, the achievable SINR of this content is scaled by $1/s^m$ in the objective function (4.98). The network resources will be biased to give more priority on this content. The max operation outside guarantees that the networks resources should be fully exploited. The transmission power vector $\mathbf{P} = [P_1, P_2, ..., P_N]^T$ indicates the maximal allowable transmission power of each eRRH [9]. As already stated and adopted in previous subsections, constraints (4.99) and (4.100) denote the power and fronthaul resource consumption at eRRH n. Note that we aim to maximize the SE of the network, thus the operational power consumption is not necessary to be considered anymore, since all eRRHs must be activated to maximize the achievable spectral efficiency.

Obviously, by introducing a scalar f, the problem above can be equivalently refor-

[9] We set the maximal allowable transmission power vector as an input variable of this problem, for the solving procedure to be introduced in this subsection later.

mulated as:

$$\mathcal{F}_{\text{wMMF}}(\mathbf{s},\mathbf{P}): \quad \max_{\{\mathbf{v}^m\}_{m=1}^M} \; f, \tag{4.101}$$

$$\text{s.t.} \quad \frac{\Gamma^m}{s^m} \geq f, \quad \forall k \in \mathcal{G}^m, \quad \forall \mathcal{G}^m, \tag{4.102}$$

$$\text{with } \Gamma^m = \frac{|\widetilde{\mathbf{h}}_k^H \mathbf{v}^m|^2}{\sigma_k^2 + \sum_{i\neq m}^M |\widetilde{\mathbf{h}}_k^H \mathbf{v}^i|^2},$$

$$\sum_{m=1}^M ||\mathbf{v}_n^m||_2^2 \leq P_n, \qquad \forall n \in \mathcal{N}, \tag{4.103}$$

$$\sum_{m=1}^M (1-c_n^{f^m}) \left|||\mathbf{v}_n^m||_2^2\right|_0 \log_2(1+\Gamma^m) \leq C_{\text{FH},n}, \qquad \forall n \in \mathcal{N}. \tag{4.104}$$

It can be easily noticed that both the objective function and the constraints of the problem above are non-convex and NP-hard. Solving it directly is difficult. However, as we are going to show, the problem can be solved in a tortuous manner: By introducing and solving a related problem, which is similar to a dual problem of the original one, some insights into the problem above can be obtained. Then together with the Bi-Section method and the solution of the introduced related problem, the solution of problem (4.101)-(4.104) can be finally reached.

The related problem is actually the transmission power minimization problem of the same network, i.e., the problem (4.21)-(4.23) and (4.25), introduced in Subsection 4.2.1.1 for the high EE oriented design. We just need to substitute the target SINR for content f^m in (4.22) with the scaling factor of it, i.e., s^m. For ease of further interpretation, the related problem is formulated as follows [10]:

$$\mathcal{P}_{\text{TX}}(\mathbf{s}) \qquad \min_{\{\mathbf{v}^m\}_{m=1}^M} \quad \sum_{m=1}^M ||\mathbf{v}^m||_2^2, \tag{4.105}$$

$$\text{s.t.} \quad \frac{\Gamma^m}{s^m} \geq 1 \quad \forall k \in \mathcal{G}^m, \quad \forall \mathcal{G}^m, \tag{4.106}$$

$$\text{with } \Gamma^m = \frac{|\widetilde{\mathbf{h}}_k^H \mathbf{v}^m|^2}{\sigma_k^2 + \sum_{i\neq m}^M |\widetilde{\mathbf{h}}_k^H \mathbf{v}^i|^2},$$

$$\sum_{m=1}^M (1-c_n^{f^m}) \left|||\mathbf{v}_n^m||_2^2\right|_0 \log_2(1+\Gamma^m) \leq C_{\text{FH},n}, \quad \forall n \in \mathcal{N}, \tag{4.107}$$

$$\sum_{m=1}^M ||\mathbf{v}_n^m||_2^2 \leq P_{\text{TX},n}^{\max}, \quad \forall n \in \mathcal{N}. \tag{4.108}$$

The solving procedure for problem (4.105)-(4.108) is shown in Alg. 2. Then the crucial question is: What is the relationship between problem $\mathcal{F}_{\text{wMMF}}(\mathbf{s},\mathbf{P})$ and $\mathcal{P}_{\text{TX}}(\mathbf{s})$? How can we solve the first problem with the help of the second one? Let $\mathbf{P} = [P_1, P_2, ..., P_N]^T$ denote the minimized transmission power of the network,

[10] $\{P_{\text{TX},n}^{\max}\}_{n=1}^N$ in this problem is predetermined and fixed.

which is the result [11] by solving problem $\mathcal{P}_{\mathrm{TX}}(\mathbf{s})$ with Alg. 2. Such a relationship is expressed as $\mathbf{P} = \mathcal{P}_{\mathrm{TX}}(\mathbf{s})$. Similarly, let f be the result of the problem $\mathcal{F}_{\mathrm{wMMF}}(\mathbf{s}, \mathbf{P})$, we have $f = \mathcal{F}_{\mathrm{wMMF}}(\mathbf{s}, \mathbf{P})$. Before we proceed to solve the wMMF problem, the following lemmas must be introduced firstly:

Lemma 1: Problem $\mathcal{F}$ and $\mathcal{P}$ are related as follows:

$$f = \mathcal{F}_{\mathrm{wMMF}}(\mathbf{s}, \mathcal{P}_{\mathrm{TX}}(f\mathbf{s})); \tag{4.109}$$

$$\mathbf{P} = \mathcal{P}_{\mathrm{TX}}(\mathcal{F}_{\mathrm{wMMF}}(\mathbf{s}, \mathbf{P})\mathbf{s}). \tag{4.110}$$

For (4.109), it denotes that for an arbitrary scalar f, and an arbitrary valid weighting coefficients vector $f\mathbf{s}$, we can definitely obtain the corresponding minimized power allocation scheme $\mathbf{P}$ resulting from solving $\mathcal{P}_{\mathrm{TX}}(f\mathbf{s})$. Then by setting $\mathbf{s}$ and $\mathbf{P}$ as the input parameter to the problem $\mathcal{F}_{\mathrm{wMMF}}$, the same value of f can be obtained by solving it. For (4.110), such a relationship can be interpreted similarly.

Proof: The contradiction is used for the proof. For equation (4.109), let $\{\mathbf{v}^{m,\mathrm{opt}}\}_{m=1}^{M}$ and $\mathbf{P}^{\mathrm{opt}} = [P_1^{\mathrm{opt}}, P_1^{\mathrm{opt}}, ..., P_N^{\mathrm{opt}}]^T$ denote the optimal beamformers and the optimal (minimized) power consumption of problem $\mathcal{P}_{\mathrm{TX}}(f\mathbf{s})$ respectively, where $f\mathbf{s}$ represents the SINR requirements. Then for problem $\mathcal{F}_{\mathrm{wMMF}}(\mathbf{s}, \mathbf{P}^{\mathrm{opt}})$, beamformers $\{\mathbf{v}^{m,\mathrm{opt}}\}_{m=1}^{M}$ must be a feasible solution with objective value f. If another feasible solution $\{\widetilde{\mathbf{v}}^{m}\}_{m=1}^{M}$ with objective value $\widetilde{f} > f$ exists, then a constant $c < 1$ must also exist, such that it can further scale down the solution, e.g., $\{c\widetilde{\mathbf{v}}^{m}\}_{m=1}^{M}$, with which the SINR requirements of $\mathcal{P}_{\mathrm{TX}}(f\mathbf{s})$, as well as the fronthaul capacity and power constraints are still fulfilled. Thus, $\{c\widetilde{\mathbf{v}}^{m}\}_{m=1}^{M}$ must result in lower power consumption than $\mathbf{P}^{\mathrm{opt}}$, which contradicts the optimality assumption of $\{\mathbf{v}^{m,\mathrm{opt}}\}_{m=1}^{M}$. Equation (4.110) can be proved similarly.

Lemma 2: For a given valid vector $\mathbf{s}$, the minimized total transmission power $\sum_{n=1}^{N} P_n^{\mathrm{opt}}$ of problem $\mathcal{P}_{\mathrm{TX}}(f\mathbf{s})$, is monotonically non-decreasing when the value of f is increased. And the value of f resultant from $\mathcal{F}_{\mathrm{wMMF}}(\mathbf{s}, \mathbf{P})$ is monotonically non-decreasing when $\sum_{n=1}^{N} P_n$ is increased.

Proof: When the value of f is increased, the SINR requirements in $\mathcal{P}_{\mathrm{TX}}(f\mathbf{s})$ become more stringent, thus the feasible set for the solution cannot be enlarged. When higher power budget is available, it can always be evenly distributed among all beamformers to increase all SINR, as long as the noise power σ_k^2 is larger than 0.

Corollary: *Lemma 1* suggests that, for a fixed scaling vector $\mathbf{s}$, the solution of

[11] The solution of the problem $\mathcal{P}_{\mathrm{TX}}$ is actually the optimized beamforming vectors $\{\mathbf{v}^{m}\}_{m=1}^{M}$, based on which the power allocation can be obtained for each eRRH, i.e., $P_n = \sum_{m=1}^{M} |\mathbf{v}_n^{m,\mathrm{opt}}||_2^2 \; \forall n \in \mathcal{N}$.

$f = \mathcal{F}_{\text{wMMF}}(\mathbf{s}, \mathbf{P})$ can always be found by solving problem $\mathbf{P}' = \mathcal{P}_{\text{TX}}(f'\mathbf{s})$ via checking different values of f', exhaustively until $\sum_{n=1}^{N} P_n = \sum_{n=1}^{N} P'_n$ satisfies. Thanks to *Lemma 2*, such an exhaustive search is not necessary, as the value of f can be located much more efficiently with the Bi-Section method. Moreover, due to the interaction between $f = \mathcal{F}_{\text{wMMF}}(\mathbf{s}, \mathbf{P})$ and $\mathcal{P}_{\text{TX}}(f'\mathbf{s})$, individual eRRH power constraints can always be satisfied due to (4.108).

Hence, based on the lemmas and corollary introduced above, together with the Bi-Section method, the solving procedure for the original wMMF problem (4.98)-(4.100), can be converted to solving several TX power minimization problems, each of them is constructed via location of the Bi-section method. In other words, problem $\mathcal{F}_{\text{wMMF}}$ is not solved directly, but it is solved by solving its related problems $\mathcal{P}_{\text{TX}}$ instead, with known algorithms. By adopting the Bi-section method, the solution of the original problem $\mathcal{F}_{\text{wMMF}}$ can be approached. The overall steps are summarized in Alg. 5:

Algorithm 5: Weighted Max-Min Fairness Optimization Steps

Initialization: Set f_L and f_U as the lower and upper bound of the searching range.
repeat
  Set $f \leftarrow (f_L + f_U)/2$. Solve the problem $\mathbf{P} = \mathcal{P}_{\text{TX}}(f\mathbf{s})$ with Alg. 2.
  if $\sum_{n=1}^{N} P_n > \sum_{n=1}^{N} P_{\text{TX},n}^{\max}$ *or the problem is infeasible* **then**
    Set $f_U \leftarrow f$.
  else
    Set $f_L \leftarrow f$.
until $f_U - f_L < \varepsilon$, *where* ε *denotes the tolerance*;
Solve the standard the SDP problem $\mathcal{P}_{\text{TX}}(f\mathbf{s})$ with Alg. 2, then perform EVD or use Gaussian randomization and scaling [KSL08] method to obtain the approximated solution $\{\mathbf{v}^m\}_{m=1}^{M}$.

Remark 1: For the lower bound and upper bound used for the Bi-Section search, the value of f_L and f_U is initialized as follows:

$$f_L = 0, \tag{4.111}$$

$$f_U = \min_{k \in \{1,2,\ldots,K\}} \frac{1}{s_{k \in \mathcal{G}^m}} \frac{||\mathbf{h}_k||_2^2 \sum_{n=1}^{N} P_{\text{TX},n}^{\max}}{\sigma_k^2}. \tag{4.112}$$

Actually, the upper bound is set to be the minimal achievable SINR, when all eRRHs contributes all their available power towards a single group, i.e., no multi-cast and interference exist in this case.

Remark 2: Note that the wMMF problem $f = \mathcal{F}_{\text{wMMF}}(\mathbf{s}, \mathbf{P})$ is always feasible, i.e., a positive maximized minimal weighted SINR f always exists for any valid $\mathbf{s}$ and $\mathbf{P}$. However, the power minimization problem $\mathbf{P} = \mathcal{P}_{\text{TX}}(\mathbf{s})$ is not necessarily feasible.

In some cases, the SINR targets $\mathbf{s}$ can not be achieved simultaneously for all multi-cast groups, with the instantaneous channel states, cache hitting status, individual fronthaul and power constraints. This can be due to many uncached contents being requested by UEs, such that the fronthaul capacities are not sufficient to deliver all of them to sufficient number of eRRHs, leading to lower array gain between eRRHs. In such cases, we also need to reduce the value of the upper bound f_U, as in Step 5 of Alg. 5.

For investigating the properties of the wMMF metric, some numerical results will be provided based on Alg. 5, but it would be better that they appear together with that of the TP-Max metric for comparison. Hence, we will firstly elaborate on the solving procedure for maximizing the multi-cast throughput in the coming subsection, which is then followed with the numerical results of both design metrics.

4.2.4 High SE oriented Design — TP-Max Metric

When we talk about the network throughput here, we mean the sum achievable rate among all requested contents [12]. As stated many times before, the achievable rate of a specific content is determined by the worst UE within the multi-cast group requesting it, i.e., for all UEs in the multi-cast group $\mathcal{G}^m$, they experience a downlink rate of $\log_2\left(1+\min_{k\in\mathcal{G}^m}\mathrm{SINR}_k\right)$. Hence, the sum multi-cast rate of the network (multi-cast throughput) can be calculated by summing up the achievable rate among all requested contents. The problem formulation for maximizing the network throughput is straightforward as follows:

$$\mathcal{T}: \max_{\{\tilde{\mathbf{v}}^m,\, p^m\}_{m=1}^M} \sum_{m=1}^{M} \log_2\left(1+\min_{k\in\mathcal{G}^m}\mathrm{SINR}_k\right), \tag{4.113}$$

$$\text{s.t.} \sum_{m=1}^{M} p^m ||\tilde{\mathbf{v}}_n^m||_2^2 \le P_{\mathrm{TX},n}^{\max}, \quad \forall n\in\mathcal{N}, \tag{4.114}$$

$$\sum_{m=1}^{M} (1-c_n^{f^m}) \left|||\tilde{\mathbf{v}}_n^m||_2^2\right|_0 \log_2\left(1+\min_{k\in\mathcal{G}^m}\mathrm{SINR}_k\right) \le C_{\mathrm{FH},n}, \quad \forall n\in\mathcal{N}, \tag{4.115}$$

with

$$\mathrm{SINR}_k = \frac{p^m |\widetilde{\mathbf{h}}_k^H \tilde{\mathbf{v}}^m|^2}{\sum_{i\neq m}^{M} p^i |\widetilde{\mathbf{h}}_k^H \tilde{\mathbf{v}}^i|^2 + \sigma_k^2}, \quad k\in\mathcal{G}^m. \tag{4.116}$$

Compared with the problem formulations in previous subsections, there is a minor modification here: We adopt the normalized aggregate beamformers $\tilde{\mathbf{v}}^m$ =

[12] For some different definitions, it usually indicates the sum achievable rate among all UEs. However, in the multi-cast scenario, like many existing works, the achievable rate of each content is considered, in order to ensure each UE requesting it can finally be served.

$[\{\tilde{\mathbf{v}}_1^m\}^H, \{\tilde{\mathbf{v}}_2^m\}^H, ..., \{\tilde{\mathbf{v}}_N^m\}^H]^H \in \mathbb{C}^{NL\times 1}$ among all eRRHs, such that $||\tilde{\mathbf{v}}^m||_2^2 = \sum_{n=1}^N ||\tilde{\mathbf{v}}_n^m||_2^2 = 1$, where $\tilde{\mathbf{v}}_n^m = [\tilde{v}_{n,1}^m, \tilde{v}_{n,2}^m, ..., \tilde{v}_{n,L}^m]^T \in \mathbb{C}^{L\times 1}$ indicates the part of the normalized beamformer constructed at eRRH n. p^m is used to denote the power allocated to content f^m for all UEs in multi-cast group $\mathcal{G}^m$, and vector $\mathbf{p} = [p^1, p^2, ..., p^M]^T$ indicate the power allocation scheme to all M multi-cast groups. Hence, the relationship between the aggregated beamformer, which is always used in previous problems, and the normalized aggregated beamformer is

$$\mathbf{v}^m = \sqrt{p^m}\tilde{\mathbf{v}}^m = \sqrt{p^m}[\{\tilde{\mathbf{v}}_1^m\}^H, \{\tilde{\mathbf{v}}_2^m\}^H, ..., \{\tilde{\mathbf{v}}_N^m\}^H]^H \in \mathbb{C}^{NL\times 1}, \forall m \in \mathcal{M}. \quad (4.117)$$

The reason to introduce such normalized beamformers with power allocation vector is, when the multi-cast throughput maximization is considered, the power allocated for each requested content shall also be directly optimized. By splitting the beamformers into the part of the power and the part of the normalized beamformers, we have the chance to directly manipulate the power allocation, as well as the beamformer directions. By optimizing both $\{\tilde{\mathbf{v}}^m, p^m\}_{m=1}^M$, the optimal eRRH cluster formulation (via computing the ℓ_0-norm of the optimized normalized beamformers) and the power allocation can be obtained to maximize the multi-cast network throughput.

In constraint (4.114), $p^m||\tilde{\mathbf{v}}_n^m||_2^2$ denotes the power at eRRH n, which is allocated to serve the multi-cast group $\mathcal{G}^m$. Hence, the LHS of (4.114) indicates the total transmission power of eRRH n, which shall not exceed its maximal allowable power. Constraint (4.115) guarantees the fronthaul connected to each eRRH can support the data streams that deliver the uncached contents.

Unfortunately, the problem above is rather difficult to solve. Although we have known how to use the SDR and the iterative ℓ_0-norm approximation method to convexify several parts of the problem, the difficulty mainly lies at (4.113) and (4.115), where the min operation makes them no longer differentiable. Hence, compared with the problem solved in [KPS12; CCO14], where only a single RRH exists in C-RAN, the multi-cast throughput maximization in the cache-enabled F-RAN is much more complicated. However, thanks to the clever heuristic ideas used there, their thoughts are extended to solve the problem here.

For a better understanding, we firstly sketch the idea of the proposed algorithm, which gives an intuitive explanation about how and why it works. The algorithm will be introduced in detail afterwards.

4.2.4.1 Basic Idea and Sketch of the Algorithm

Due to the interaction between two types of optimization variables, i.e., the normalized multi-cast beamformers $\{\tilde{\mathbf{v}}^m\}_{m=1}^M$ and the power allocation scheme $\mathbf{p}$ in

problem (4.113)-(4.115), a simultaneous optimization of these two types of variables is difficult. Thus, the solving procedure is designed to perform in an **alternating** way: Each alternating step consists of two sub-steps, i.e., the **Re-Design sub-step** and the **Re-Allocation sub-step**. In each sub-step, one specific variable type is fixed and the other is to be optimized. Then in the next sub-step, the one is fixed, which is just optimized, and the other variable type which is fixed in the previous sub-step is to be optimized.

1. At the t-th alternating step, the aggregated multi-cast beamformers $\{\mathbf{v}^{m(t-1)}\}_{m=1}^{M} = \left\{\sqrt{p^{m(t-1)}} \cdot \tilde{\mathbf{v}}^{m(t-1)}\right\}_{m=1}^{M}$ are supposed to be known from the last alternating step. Hence, the current achievable SINRs $\{\Gamma^{m(t-1)}\}_{m=1}^{M} = \left\{\min_{k\in\mathcal{G}^m} \mathrm{SINR}_k^{(t-1)}\right\}_{m=1}^{M}$ of all contents can be computed according to (4.116).

2. **Re-Design sub-step**:. Now $\boldsymbol{\Gamma}^{(t-1)} = \{\Gamma^{m(t-1)}\}_{m=1}^{M}$ is set to be the SINR target, then a related power minimization problem $\mathcal{P}^{(t)}(\boldsymbol{\Gamma}^{(t-1)})$[13] is constructed and solved to optimize and obtain new multi-cast beamformers $\{\mathbf{v}^m\}_{m=1}^{M}$ (Re-Design), such that the same multi-cast sum rate (network throughput (4.113)) can be achieved [14], but with **less power consumption**. It worth to mention that although the transmission power is to be minimized here, such a target is achieved by optimizing the beamforming vectors. Thus, this sub-step redesigns the beamformers, which will be used in the next sub-step. The reduction of the power consumption in this sub-step is always possible for the multi-cast case: Note that the achievable rate of each multi-cast group is limited by SINR of the UE with the worst channel conditions in this group, thus the actual SINR at other UEs of the same group might be much higher. Hence, by re-designing the beamformers via solving problem $\mathcal{P}^{(t)}(\boldsymbol{\Gamma}^{(t-1)})$, such *useless* higher SINRs at side of other UEs in this multi-cast group can be reduced to the same level of the worst UE. At the same time, the multi-cast throughput can still stay unchanged. In summary, with the re-designed beamformers $\{\mathbf{v}^m\}_{m=1}^{M}$, the same network performance in terms of the multi-cast throughput can be achieved, but with less power consumption, compared with the scheme $\{\mathbf{v}^{m(t-1)}\}_{m=1}^{M}$ from the last alternating step.

3. Now we see that the newly generated $\{\mathbf{v}^m\}_{m=1}^{M} = \{\sqrt{p^m} \cdot \tilde{\mathbf{v}}^m\}_{m=1}^{M}$ can save power compared with $\{\mathbf{v}^{m(t-1)}\}_{m=1}^{M}$, without the performance loss in terms of the multi-cast throughput. Hence, the beamformers of alternating step t can be updated as $\{\mathbf{v}^{m(t)}\}_{m=1}^{M} = \{\mathbf{v}^m\}_{m=1}^{M}$, and the corresponding normalized beamformers are updated by computing $\left\{\mathbf{v}^{m(t)} / \sqrt{||\mathbf{v}^{m(t)}||_2^2}\right\}_{m=1}^{M}$.

[13] This is actually the problem (4.21)-(4.25)

[14] We have set the achieved SINR for realizing the multi-cast throughput from the last sub-step as the new SINR targets in this problem, thus the new problem is definitely feasible and solvable.

4. **Re-Allocation sub-step**: As some power is saved in the previous sub-step, some extra power budget is now available. If the extra power budget can be somehow re-distributed to simultaneously increase all SINRs, the multi-cast throughput can be further increased. Thus, in this sub-step, all available power is to be re-allocated among eRRHs. We will show that such a goal can be achieved by using the sub-gradient method [BV04]. Let $\{p^m\}_{m=1}^M$ be the resultant power allocation of this method, the power allocation scheme is updated as $\{p^{m(t)}\}_{m=1}^M = \{p^m\}_{m=1}^M$. Although the new power allocation further increases the multi-cast throughput, it generates some *useless* higher SINRs at some UEs in each multi-cast group again, which cannot contribute to the increase of the multi-cast throughput, due to the worst UE in this group. Then the $(t+1)$-th alternating step starts, and the Re-Design sub-step will update the beamformers to save power.

After the general introduction of the basic idea, we now deep into each sub-step.

4.2.4.2 Beamformer Updates via the Re-Design Sub-step

As discussed above, in the Re-Design sub-step, a power minimization problem is to be solved for re-designing the beamformers, such that the same network performance in terms of the multi-cast throughput can still be achieved, but with less transmission power. Obviously, it is just the problem that we have solved in Subsection 4.2.1.1, but with $\mathbf{\Gamma}$ computed from the last alternating step as the input parameter [15]. The power minimization problem for t-th alternating step is:

$$\mathcal{P}^{(t)}(\mathbf{\Gamma}^{(t-1)}): \quad \min_{\{\mathbf{v}^m\}_{m=1}^M} \quad \sum_{m=1}^M ||\mathbf{v}^m||_2^2, \tag{4.118}$$

$$\text{s.t.} \quad \text{SINR}_k^{\text{hard}} \geq \Gamma^{m(t-1)}, \quad \forall k \in \mathcal{G}^m, \quad \forall \mathcal{G}^m, \tag{4.119}$$

$$\sum_{m=1}^M (1 - c_n^{f^m}) |||\mathbf{v}_n^m||_2^2|_0 \log_2\left(1 + \Gamma^{m(t-1)}\right) \leq C_{\text{FH},n}, \quad \forall n \in \mathcal{N}, \tag{4.120}$$

$$\sum_{m=1}^M ||\mathbf{v}_n^m||_2^2 \leq P_{\text{TX},n}^{\max}, \quad \forall n \in \mathcal{N}. \tag{4.121}$$

After solving the problem above with Alg. 2 introduced in Subsection 4.2.1.1, the resultant re-designed beamformers are set to $\{\mathbf{v}^{m(t)}\}_{m=1}^M$, which is the re-designed beamformers of the t-th iteration, and will be used in the Re-Allocation sub-step for optimizing the power allocation.

4.2.4.3 Power Allocation via the Re-Allocation Sub-step

The algorithm for the power allocation is something fresh new! As in the original problem $\mathcal{T}$ (4.113)-(4.115), when the beamformers are known,

[15] We still take the hard transfer mode with dedicated fronthaul as the example.

the multi-cast throughput depends solely on the power allocation scheme $\mathbf{p}$. Hence, in this sub-step, the normalized aggregated beamformers $\{\tilde{\mathbf{v}}^{m(t)}\}_{m=1}^{M} = \left\{\mathbf{v}^{m(t)}/\sqrt{||\mathbf{v}^{m(t)}||_2^2}\right\}_{m=1}^{M}$ are fixed, which are obtained from the Re-Design sub-step, and the power budget saved from the previous sub-step will be redistributed, so as to further increase the throughput. The power allocation problem for the t-th alternating step can be formulated as follows:

$$\mathcal{R}^{(t)}(\{\tilde{\mathbf{v}}^{m(t)}\}_{m=1}^{M}): \quad \max_{\mathbf{p}} \quad \sum_{m=1}^{M} \log_2\left(1+\min_{k\in\mathcal{G}^m}\mathrm{SINR}_k(\mathbf{p})\right), \tag{4.122}$$

$$\text{s.t.} \sum_{m=1}^{M} p^m ||\tilde{\mathbf{v}}_n^{m(t)}||_2^2 \le P_{\mathrm{TX},n}^{\max}, \quad \forall n\in\mathcal{N}, \tag{4.123}$$

$$\sum_{m=1}^{M}(1-c_n^{f^m})\left|||\tilde{\mathbf{v}}_n^{m(t)}||_2^2\right|_0 \log_2\left(1+\min_{k\in\mathcal{G}^m}\mathrm{SINR}_k(\mathbf{p})\right) \le C_{\mathrm{FH},n}, \quad \forall n\in\mathcal{N}. \tag{4.124}$$

As seen from (4.116), the achievable SINR for each UE is a function of the power allocation scheme $\mathbf{p}$. Obviously, as the aggregated normalized beamformers resulting from the last sub-step are the input parameters and fixed, constraint (4.123) is linear with respect to $\mathbf{p}$, and both objective (4.122) and constraint (4.124) are linear with respect to $\log_2\left(1+\min_{k\in\mathcal{G}^m}\{\mathrm{SINR}_k(\mathbf{p})\}\right)$. In order to deal with the non-convex and non-differentiable term $\log_2\left(1+\min_{k\in\mathcal{G}^m}\{\mathrm{SINR}_k(\mathbf{p})\}\right)$, we firstly introduce the following Proposition, with which this *nasty* term can be approximated and convexified.

Proposition [KPS12]: Let $\omega^m > 0\ \forall m$ be constants, and vector $\mathbf{s} = [s^1, s^2, ..., s^M]^T$ be parameters input to the function, then

$$f(\mathbf{s}) = \sum_{m=1}^{M} \omega^m \psi\left(\min_{k\in\mathcal{G}^m}\mathrm{SINR}_k(e^{\mathbf{s}})\right) \tag{4.125}$$

is a non-differentiable convex function of $\mathbf{s}$, as long as

1. ψ is continuous differential and strictly decreasing;
2. the inversion of $-\psi$, i.e., $(-\psi)^{-1}$ is log-convex [16].

As this proposition has been proved in [KPS12], we just adopt it here without proof. For ease of further discussion, $R(\mathbf{p})$ is used to denote the multi-cast throughput under the specific power allocation scheme $\mathbf{p}$, i.e.,

$$R(\mathbf{p}) = \sum_{m=1}^{M} \log_2\left(1+\min_{k\in\mathcal{G}^m}\mathrm{SINR}_k(\mathbf{p})\right). \tag{4.126}$$

[16] A function f(x) is said to be log-convex on interval $[a,b]$ when $f(x) > 0$ and $\ln f(x)$ is convex on $[a,b]$.

By comparing (4.125) and (4.126), it can be observed that if we set $\mathbf{s} = \ln \mathbf{p}$ and $\psi(x) = -\log_2(x)$, the multi-cast throughput maximization problem via power allocation $\mathbf{p}$, i.e. $\max_{\mathbf{p}} R(\mathbf{p})$, can be well approximated by searching for $\mathbf{s}$, which minimizes $f(\mathbf{s})$. Such an approximation is more precise in the **high SINR regime**. Thanks to the F-RAN architecture, where a BBU pool and multiple eRRHs can form a distributed MIMO structure, much more concentrated beams are possible. Thus, much higher SINRs than the single BS scenario [KPS12] can be achieved more probable. Thus, such an approximation is particular suitable for the F-RAN scenario.

As $f(\mathbf{s})$ is non-differentiable convex, the sub-gradient method shall be exploited. Once the optimal $\mathbf{s}^{\text{opt}}$ is obtained, the optimal power allocation can be computed via $\mathbf{p}^{\text{opt}} = \exp(\mathbf{s}^{\text{opt}})$. However, the power and the fronthaul capacity constraints make the problem much more complicated. By setting $\{\tilde{\mathbf{v}}^m\}_{m=1}^M$ as the fixed input parameters, and adopting the proposition introduced above, the original power allocation problem (4.122)-(4.124) can be approximated as follows:

$$\mathcal{R}(\{\tilde{\mathbf{v}}^{m(t)}\}_{m=1}^M) : \min_{\mathbf{s}} \; f(\mathbf{s}), \tag{4.127}$$

$$\text{s.t.} \sum_{m=1}^{M} v_n^{m(t)} p^m - P_{\text{TX},n}^{\max} \leq 0, \quad \forall n \in \mathcal{N}, \tag{4.128}$$

$$h_n(\mathbf{s}) - C_{\text{FH},n} \leq 0, \quad \forall n \in \mathcal{N}, \tag{4.129}$$

$$\text{where } \psi(x) = -\log_2(x), \; \mathbf{p} = \exp(\mathbf{s}), \tag{4.130}$$

$$f(\mathbf{s}) = \sum_{m=1}^{M} \psi\left(\min_{k \in \mathcal{G}^m} \text{SINR}_k(e^{\mathbf{s}})\right), \tag{4.131}$$

$$h_n(\mathbf{s}) = \sum_{m=1}^{M} \omega_n^{m(t)} \psi\left(\min_{k \in \mathcal{G}^m} \text{SINR}_k(e^{\mathbf{s}})\right), \quad \forall n \in \mathcal{N}, \tag{4.132}$$

$$v_n^{m(t)} = ||\tilde{\mathbf{v}}_n^{m(t)}||_2^2 \geq 0, \quad \forall m \in \mathcal{M}, \forall n \in \mathcal{N}, \tag{4.133}$$

$$\omega_n^{m(t)} = -(1 - c_n^{f^m}) \left| ||\tilde{\mathbf{v}}_n^{m(t)}||_2^2 \right|_0 \leq 0, \quad \forall m \in \mathcal{M}, \forall n \in \mathcal{N}. \tag{4.134}$$

Constraints (4.128) and (4.129) result from (4.123) and (4.124), which ensure that the solution $\mathbf{p}^{\text{opt}}$, as well as the resultant fronthaul requirements can be supported at each eRRH. Coefficients $v_n^{m(t)}$ and $\omega_n^{m(t)}$ are constants computed via the beamformers, which are obtained in the Re-Design sub-step, and are fixed here.

According to the proposition, the objective (4.127) is a non-differentiable convex function of $\mathbf{s}$. Constraints (4.128) are linear functions of $\mathbf{p}$ and thus form a convex set. However, constraints (4.129) are concave due to $\omega_n^{m(t)} \leq 0$. The sub-gradient method (with general convex constraints) cannot be applied directly. Fortunately, by reviewing $f(\mathbf{s})$ and $h_n(\mathbf{s})$, it can be observed that they have the same structure but with different coefficients. Therefore, when the LHS of (4.129) equals 0 with a specific $\mathbf{s}$, i.e., the fronthaul resources have been completely exhausted, the multi-cast throughput in (4.127) cannot be increased further via re-distributing the power,

as it will inevitably lead to the violation of (4.129). When there are still available fronthaul resources, i.e., the LHS of (4.129) is smaller than 0, the sub-gradient evolution can be executed further on (4.127) as long as the step size is small enough, until the constraints (4.129) are violated.

To fulfill the linear constraints (4.128), the projected sub-gradient [BV04] can be exploited, which is also used in [CCO14] to fulfill the per-antenna power constraints. More details of this method has already been introduced in Subsection 2.3.3. The convex set $\mathcal{C}\left(\{\tilde{\mathbf{v}}_n^{m(t)}\}_{m=1}^M\right)$ defined by (4.128) and (4.133) can be expressed as:

$$\mathcal{C}\left(\{\tilde{\mathbf{v}}_n^{m(t)}\}_{m=1}^M\right) = \left\{\mathbf{p} \in \mathbb{R}_+^{M\times 1} | \boldsymbol{\nu}^{(t)}\mathbf{p} \leq \mathbf{P}_{\text{TX}}^{\max}\right\}, \quad \text{where } \boldsymbol{\nu}^{(t)} \in \mathbb{R}_+^{N\times M} \text{ with } (n,m)\text{-th element as } \nu_n^{m(t)}. \tag{4.135}$$

Here vector $\mathbf{P}_{\text{TX}}^{\max} = [P_{\text{TX},1}^{\max}, P_{\text{TX},2}^{\max}, ..., P_{\text{TX},N}^{\max}]^T$ indicates the maximal allowable transmission power of each eRRH. Now we summarize the projected sub-gradient searching steps for problem $\mathcal{R}(\{\tilde{\mathbf{v}}_n^{m(t)}\}_{m=1}^M)$ as follows:

1. Perform the ℓ-th sub-gradient evolution

$$\widetilde{\mathbf{s}} = \mathbf{s}(\ell) - \Delta \cdot \mathbf{g}(\ell), \tag{4.136}$$

where $\mathbf{g}(\ell) = [g^1(\ell), g^2(\ell), ..., g^M(\ell)]^T$ denotes the sub-gradients of $f(\mathbf{s}(\ell))$ at $\mathbf{s}(\ell)$, which are expressed in (4.137), in which the predetermined factor Δ denotes the step size.

$$g^m(\ell) = \exp(s^m(\ell))\left(q_{\kappa^m}(\mathbf{s}(\ell)) - \sum_{\substack{i\neq m\\ i\in\mathcal{M}}} \frac{|\mathbf{h}_{\kappa^i}^H\tilde{\mathbf{v}}^{m(t)}|^2 \text{SINR}_{\kappa^i}(e^{\mathbf{s}(\ell)}) q_{\kappa^i}(\mathbf{s}(\ell))}{|\mathbf{h}_{\kappa^i}^H\tilde{\mathbf{v}}^{i(t)}|^2}\right), \tag{4.137}$$

$$\text{with } \kappa^m = \arg\min_{k\in\mathcal{G}^m} \text{SINR}_k(e^{\mathbf{s}(\ell)}),\ q_{\kappa^m}(\mathbf{s}(\ell)) = \frac{\psi'\left(\text{SINR}_{\kappa^m}(e^{\mathbf{s}(\ell)})\right)|\mathbf{h}_{\kappa^m}^H\tilde{\mathbf{v}}^{m(t)}|^2}{\sum_{\substack{i\neq m\\ i\in\mathcal{M}}} e^{s_i(\ell)}|\mathbf{h}_{\kappa^m}^H\tilde{\mathbf{v}}^{i(t)}|^2 + \sigma_{\kappa^m}^2}, \quad \forall m \in \mathcal{M}. \tag{4.138}$$

2. Check if $h_n(\widetilde{\mathbf{s}}) - C_{\text{FH},n} \leq 0,\ \forall n \in \mathcal{N}$ are fulfilled. If yes, compute $\widetilde{\mathbf{p}} = \exp(\widetilde{\mathbf{s}})$, and perform the Euclidean projection [BV04] to the convex set $\mathcal{C}\left(\{\tilde{\mathbf{v}}_n^{m(t)}\}_{m=1}^M\right)$, then update $\mathbf{s}(\ell+1)$ and continue the sub-gradient search, i.e.,

$$\mathbf{s}(\ell+1) = \ln\left(\Pi_{\mathcal{C}}(\widetilde{\mathbf{p}})\right). \tag{4.139}$$

where $\Pi_{\mathcal{C}}$ denotes the Euclidean projection to convex set $\mathcal{C}$. If it is not fulfilled, set $\mathbf{p}^{\text{opt}} = \mathbf{p}(\ell) = \exp(\mathbf{s}(\ell))$ be the solution, then terminate the projected sub-gradient search.

3. Perform the two steps above iteratively until convergence or reaching the maximal step limit, output $\mathbf{p}^{\text{opt}} = \exp(\mathbf{s}(\text{last}))$.

4.2.4.4 The Alternating Optimization Procedure

As illustrated previously, in order to maximize the multi-cast throughput, the aggregated beamformers and the power allocation scheme are optimized alternatively. Hence, after the elaboration of each optimization step, the overall alternating steps are summarized in Alg. 6:

Algorithm 6: Alternating Steps for Multi-cast Throughput Maximization

1 **Initialization**: Set $\tilde{\mathbf{v}}^{m(0)} \leftarrow \sqrt{\frac{\sum_{n=1}^{N} P_{\text{TX},n}^{\max}}{LM}} \mathbf{1}_{L\times 1}$, compute $\Gamma^{m(0)} = \min_{k\in\mathcal{G}^m} \{\text{SINR}_k\}\ \forall m$ according to (4.116). Check if $\sum_{m=1}^{M} \log_2\left(1+\Gamma^{m(0)}\right) \leq C_{\text{FH},n}$ are fulfilled. If not, further scale down and update all initial beamformers until they can be fulfilled. Set $t \leftarrow 1$.

2 **repeat**

3 Set $\mathbf{\Gamma}^{(t-1)} = [\Gamma^{1(t-1)}, \Gamma^{2(t-1)}, ..., \Gamma^{m(t-1)}]^T$ be the SINR target, then construct the power minimization problem $\mathcal{P}^{(t)}(\mathbf{\Gamma}^{(t-1)})$ according to (4.118)-(4.121) and solve it using Alg. 2. Let $\{\mathbf{v}^m\}_{m=1}^M$ be the solution.

4 Update the normalized beamformers: $\{\tilde{\mathbf{v}}^{m(t)}\}_{m=1}^M \leftarrow \left\{\mathbf{v}^m / \sqrt{||\mathbf{v}^m||_2^2}\right\}_{m=1}^M$.

5 Fix $\{\tilde{\mathbf{v}}^{m(t)}\}_{m=1}^M$ and construct the power re-distribution problem $\mathcal{R}\left(\{\tilde{\mathbf{v}}^{m(t)}\}_{m=1}^M\right)$ according to (4.127)-(4.134), perform the projected sub-gradient search according to the descriptions (4.135)-(4.139). Let $\mathbf{p}$ be the solution.

6 Update the power allocation $\{p^{m(t)}\}_{m=1}^M = \{p^m\}_{m=1}^M$, as well as the beamformers $\{\mathbf{v}^{m(t)}\}_{m=1}^M = \{\sqrt{p^{m(t)}} \cdot \tilde{\mathbf{v}}^{m(t)}\}_{m=1}^M$.

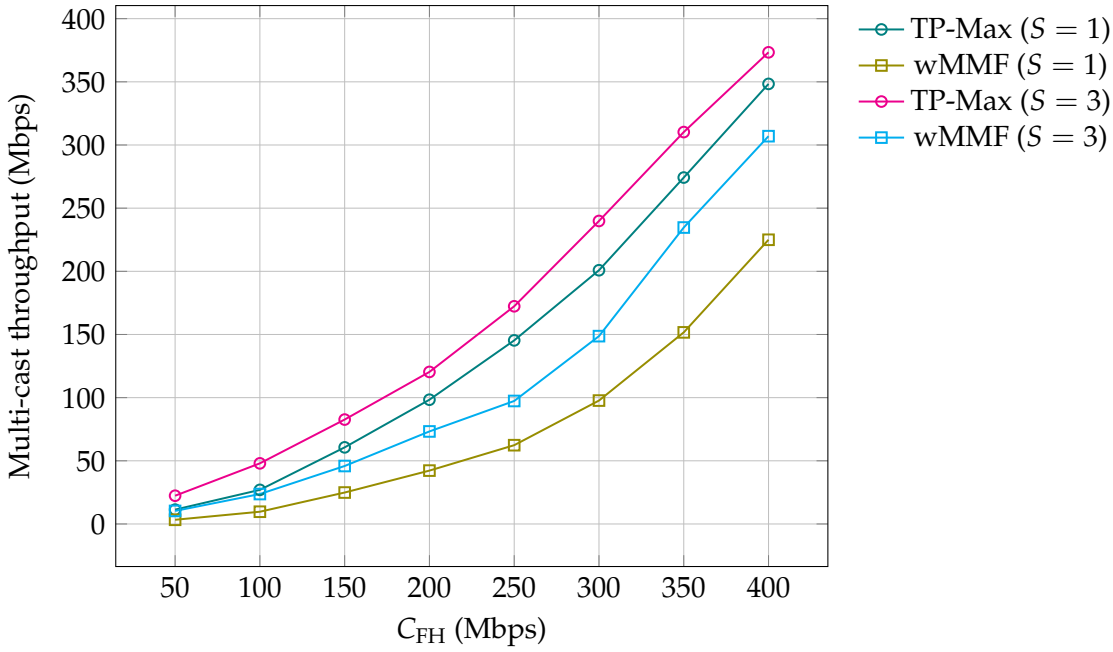
7 Compute the newly achieved SINRs $\mathbf{\Gamma}^{(t)}$ based on (4.116).

8 Set $t \leftarrow t+1$.

9 **until** *Convergence or reaching max iteration number*;

Convergence Analysis: The convergence of Alg. 6 is guaranteed: The power minimization problem $\mathcal{P}^{(t)}(\mathbf{\Gamma}^{(t-1)})$ solved in Step 3 is always feasible, since the target SINR $\mathbf{\Gamma}^{(t-1)}$ is computed based on the newly generated normalized beamformers and the power allocation scheme from the last alternating step (see Step 7), and such a design fulfills the fronthaul resource and individual power constraints due to the operations done in the Euclidean projection steps (see Step 5). As previously stated, the main purpose of $\mathcal{P}^{(t)}(\mathbf{\Gamma}^{(t-1)})$ is to reduce some *useless* high SINRs at some UEs in each multi-cast group while keep the multi-cast throughput unchanged, via the re-design of beamformers. Hence, the resultant network power consumption must be at least not higher. If some power can be saved, the re-distribution of the

Figure 4.30: The multi-cast throughput obtained for the TP-Max metric and the wMMF metric.



power must at least not lower multi-cast throughput, than the previous iteration. Therefore, such alternating procedures will definitely converge.

4.2.5 Numerical Results for wMMF Metric and TP-Max Metric

In this subsection, some numerical results of the proposed algorithms will be provided via simulation, for both wMMF and TP-Max. The same simulation environment, as well as the simulation parameters are adopted as before, unless otherwise stated. The results are based on averaging the outcome of 300 independent realizations.

In Fig. 4.30, the averaged multi-cast throughput is plotted for two metrics of high SE: wMMF resultant from Alg. 5, and TP-Max via Alg. 6. In this simulation all weight coefficients are set to be 1, i.e., each multi-cast group has the same priority. By comparing the multi-cast throughput of these two algorithms, not surprisingly, TP-Max is always higher than wMMF, as wMMF tries to balance the QoS difference between different UEs. The network might consume lots of resources to counteract the channel conditions of the *bad* UEs. However, we see that when network resources become more abundant, i.e., either more fronthaul capacity resources, or larger cache memory sizes are available, the difference between these two metrics decreases. Here we pick up some representative network configurations, under which the ratios of the throughput achieved by the wMMF metric, to that achieved by the TP-Max metric are documented. The results are listed in Table. 4.2.

Table 4.2: The ratio of the achieved multi-cast throughput: wMMF/TP-Max.

S \ C_{FH}	50 Mbps	200 Mbps	400 Mbps
1	29.2%	43.0%	64.6%
3	46.2%	60.9%	82.2%

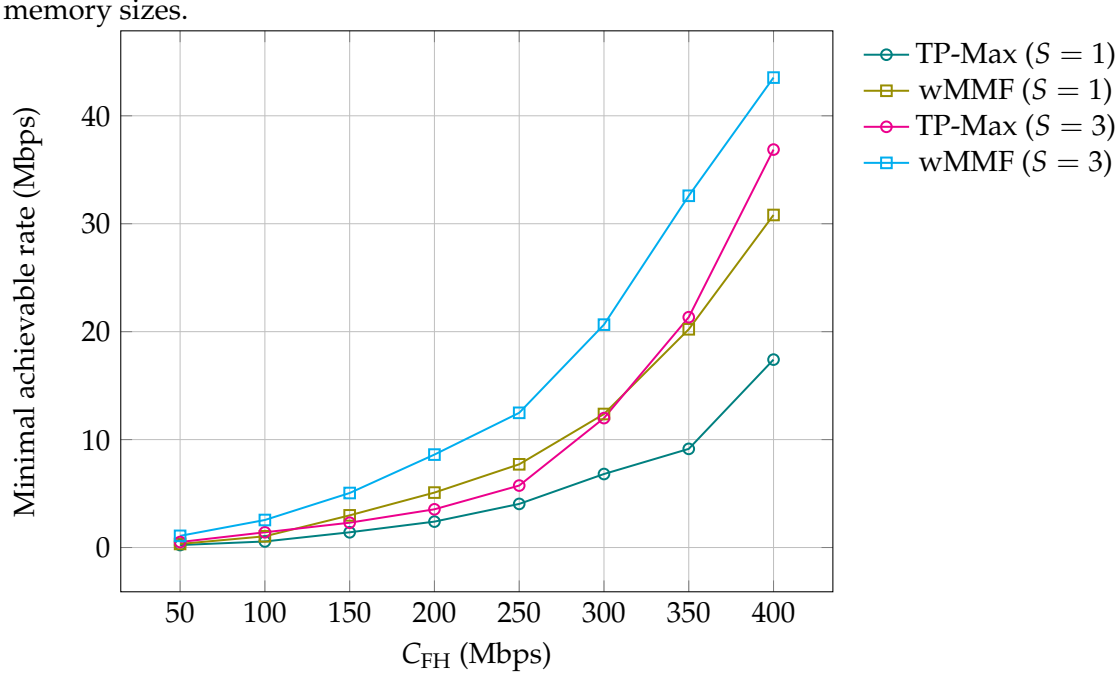
It can be observed from the table, the gap between wMMF and TP-Max becomes smaller as the network resources becomes more abundant: When cache memory has only size 1 and the fronthaul capacity is only 50 Mbps, the wMMF metric can only achieve 29.2% multi-cast throughput of the TP-Max metric. When it goes to $S = 3$ and $C_{\text{FH}} = 400$ Mbps, such a ratio can achieve 82.2%. The rationale is straightforward: When the network resources are limited, there are less eRRHs in each cluster for serving a specific multi-cast group. Hence, the wMMF oriented design is harder to combat against the bad channel conditions of some UEs. In this case, most resources have to be prioritized only for these *bad* UEs, although the resources are already rather limited! However, in the TP-Max oriented design, such *bad* UEs might be even skipped, most resources are prioritized to good UEs for improving the multi-cast throughput. Hence, the gap between these two metrics is rather large. When the network resources become more abundant, each eRRH can participate in more clusters to serve more UEs. In this case, the aggregated array gain is large enough, such that it can counteract the bad channel conditions easily. Hence, the gap between these two metrics becomes smaller.

However, when we inspect the achievable rate of the UE with the worst channel conditions, it is another story. In Fig. 4.31, the averaged minimal achieved rates for these two metrics are compared. For the same network configuration, the wMMF oriented design always outperforms the TP-Max oriented design. It is also observed that when more network resources are available, their performance become closer. The reason is the same as previously stated. By comparing Fig. 4.30 and Fig. 4.31, we can conclude that both design metrics have their own significance, depending on different service objectives. With the proposed algorithms, the BBU pool has the ability to dynamically change the performance target.

It is also worth to mention that by increasing the cache memory size, the SE performance of both metrics can also be improved significantly. Similarly, this is also due to more cooperation between eRRHs becomes more probable. Hence, besides EE, the cache is also a cheap and low-cost solution when high SE is the design target.

Before we close this subsection, we would like to illustrate the simulated convergence behaviour of Alg. 6, i.e., how such alternating steps behave. We set $P_{\text{TX},n}^{\max} = 1$ W (0 dBW) $\forall n \in \mathcal{N}$ and total $C_{\text{FH}} = 400$Mbps, each fronthaul is assumed to has the same capacity. Moreover, two different sub-gradient step sizes Δ are selected

Figure 4.31: The achieved rate for the UE with the worst channel condition, for the TP-Max metric and the wMMF metric with different fronthaul capacities and cache memory sizes.



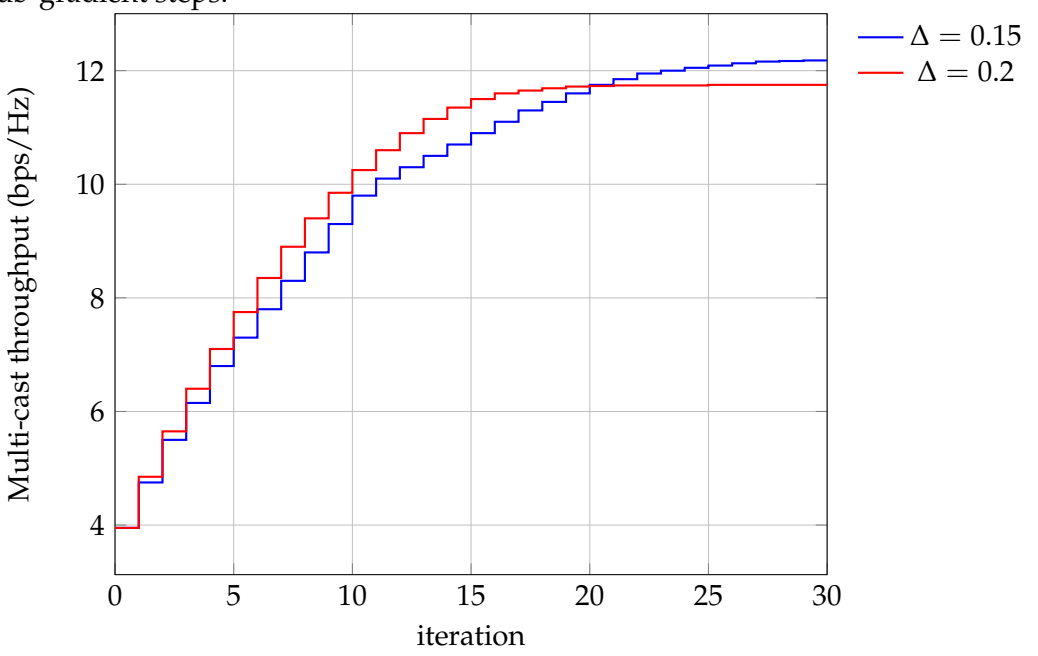
to compare the outcome. Then Alg. 6 is executed step by step, and the multi-cast throughput in each alternating step is recorded, based on the normalized beamformers and the power allocation scheme computed in the two sub-steps. The results are shown in Fig. 4.32.

The results validate our convergence analysis of Alg. 6. In each alternating step, the multi-cast throughput indeed not decreases. It converges after about 19 iterations when $\Delta = 0.2$, and about 27 iterations when $\Delta = 0.15$. While in this example, within ten iterations, 90% of the optimal performance for both cases can be achieved. By reducing the value of Δ, the algorithm requires more iterations to converge, but the resultant multi-cast throughput when it converges becomes higher. Hence, the selection of Δ reflects a trade-off between precision and complexity .

4.3 Robust Design based on Inaccurate CSI

Up to now, we have intensively discussed the optimal design of the cache-enabled F-RAN. For both high EE and SE oriented design, several algorithms have been introduced for the optimization. The numerical results demonstrated not only the effectiveness and correctness of them, but also the benefits of introducing cache modules at the network edge to perform the fog computing. However, all of the discussions and results above assume perfect CSI available at the BBU pool. In practice, the downlink CSI is actually estimated by UEs and feed back to the BBU

Figure 4.32: The convergence behaviour of the multi-cast throughput for different sub-gradient steps.



pool via the PUCCH. Therefore, the distortion is inevitable. Then some questions arise naturally: Is it possible to guarantee the network performance in the presence of only inaccurate CSI? If so, how shall the BBU pool deal with the inaccuracy in the cache-enabled F-RAN? Are the proposed algorithms in previous subsections extendable to such scenarios? Fortunately, the answer is yes, and this problem will be addressed in this subsection. Similar to the scenarios with perfect CSI, both high EE and SE oriented design under inaccurate CSI will be discussed.

4.3.1 High EE oriented Robust Design

At first, the minimization of the total power consumption is to be investigated, i.e., both transmission power and all other operational power are considered, with inaccurate CSI knowledge at the BBU pool. The power model and the inaccurate CSI model have already been introduced in Subsection 4.1.3 and Subsection 4.1.7 respectively. Moreover, the expressions of the achievable *effective* SINR for each UE are also given for the hard transfer mode (4.19), and the soft transfer mode (4.20). For ease of further illustration, we list them here again:

$$e\mathrm{SINR}_k^{\mathrm{hard}}\left(\mathbf{e}_k^{\mathrm{CSI}}\right) = \frac{\left|\mathbf{h}_k^H \mathbf{v}^m\right|^2}{\left|\mathbf{e}_k^{\mathrm{CSI}\,H} \mathbf{v}^m\right|^2 + \sum_{i\neq m}^{M} \left|\left(\mathbf{h}_k^H + \mathbf{e}_k^{\mathrm{CSI}\,H}\right)\mathbf{v}^i\right|^2 + \sigma_k^2}, \tag{4.140}$$

$$e\mathrm{SINR}_k^{\mathrm{soft}}\left(\mathbf{e}_k^{\mathrm{CSI}}\right) = \frac{\left|\mathbf{h}_k^H \mathbf{w}^m\right|^2}{\left|\left(\mathbf{h}_k^H + \mathbf{e}_k^{\mathrm{CSI}\,H}\right)\mathbf{q}\right|^2 + \left|\mathbf{e}_k^{\mathrm{CSI}\,H} \mathbf{w}^m\right|^2 + \sum_{i\neq m}^{M} \left|\left(\mathbf{h}_k^H + \mathbf{e}_k^{\mathrm{CSI}\,H}\right)\mathbf{w}^i\right|^2 + \sigma_k^2}. \tag{4.141}$$

Obviously, the achievable effective SINR at each UE depends on the CSI error vector $\{\mathbf{e}_k^{\text{CSI}}\}_{k=1}^K$, which is unknown. The BBU pool only knows that such errors are bounded within a sphere with radius ϵ_k, with the probability of at least $1-\delta_k$, as shown in (4.17) and (4.18). Since only inaccurate $\{\mathbf{h}_k\}_{k=1}^K$ CSI knowledge and the value of $\{\epsilon_k\}_{k=1}^K$ are available, the network has to be somehow optimized based on such inaccuracies, while the resultant network can still guarantee the QoS of each UE, as well as fulfill its constraints in terms of power and fronthaul capacity. In other words, the network must be *robust* to the uncertainty of the CSI knowledge.

Hence, we can formulate the problems to be solved as follows, for both hard and soft transfer mode:

$$\mathcal{P}^{\text{Hard}}_{\text{Inaccurate CSI}}: \min_{\{\mathbf{v}^m\}_{m=1}^M} \frac{1}{\xi}\left(\sum_{m=1}^M ||\mathbf{v}^m||_2^2\right) + \sum_{n=1}^N P_{\text{o}} \Big| \sum_{m=1}^M ||\mathbf{v}_n^m||_2^2 \Big|_0 + \sum_{n=1}^N P_{\text{sleep}} \left(1 - \Big| \sum_{m=1}^M ||\mathbf{v}_n^m||_2^2 \Big|_0\right), \tag{4.142}$$

$$\text{s.t.} \min_{||\mathbf{e}_k^{\text{CSI}}||_2^2 \le \epsilon_k^2} e\text{SINR}_k^{\text{hard}}(\mathbf{e}_k^{\text{CSI}}) \ge \Gamma^m, \quad \forall k \in \mathcal{G}^m, \quad \forall m \in \mathcal{M}, \tag{4.143}$$

$$\sum_{m=1}^M (1 - c_n^{f^m}) \Big| ||\mathbf{v}_n^m||_2^2 \Big|_0 \log_2(1+\Gamma^m) \le C_{\text{FH},n} \quad \forall n \in \mathcal{N} \qquad \textit{dedicated}, \tag{4.144}$$

$$\sum_{n=1}^N \sum_{m=1}^M (1 - c_n^{f^m}) \Big| ||\mathbf{v}_n^m||_2^2 \Big|_0 \log_2(1+\Gamma^m) \le C_{\text{FH}} \qquad \textit{non-dedicated}, \tag{4.145}$$

$$\sum_{m=1}^M ||\mathbf{v}_n^m||_2^2 \le P_{\text{TX},n}^{\max}, \quad \forall n \in \mathcal{N}. \tag{4.146}$$

$$\mathcal{P}^{\text{Soft}}_{\text{Inaccurate CSI}}: \min_{\{\mathbf{w}^m\}_{m=1}^M, \mathbf{q}} \frac{1}{\xi}\left(\sum_{m=1}^{M}||\mathbf{w}^m||_2^2+\sum_{n=1}^{N}||\mathbf{q}_n||_2^2\right)+\sum_{n=1}^{N}P_{\text{o}}\Big|\sum_{m=1}^{M}||\mathbf{w}_n^m||_2^2\Big|_0$$
$$+\sum_{n=1}^{N}P_{\text{sleep}}\left(1-\Big|\sum_{m=1}^{M}||\mathbf{w}_n^m||_2^2\Big|_0\right), \tag{4.147}$$
$$\text{s.t.} \min_{||\mathbf{e}_k^{\text{CSI}}||_2^2\le\epsilon_k^2} e\text{SINR}_k^{\text{soft}}(\mathbf{e}_k^{\text{CSI}})\ge\Gamma^m, \quad \forall k\in\mathcal{G}^m, \quad \forall m\in\mathcal{M}, \tag{4.148}$$
$$\sum_{l=1}^{L}\log_2\left(1+\frac{\sum_{m=1}^{M}(1-c_n^{f^m})|w_{n,l}^m|^2}{q_{n,l}^2}\right)\le C_{\text{FH},n} \quad \forall n\in\mathcal{N} \qquad \textit{dedicated}, \tag{4.149}$$
$$\sum_{n=1}^{N}\sum_{l=1}^{L}\log_2\left(1+\frac{\sum_{m=1}^{M}(1-c_n^{f^m})|w_{n,l}^m|^2}{q_{n,l}^2}\right)\le C_{\text{FH}} \qquad \textit{non-dedicated}, \tag{4.150}$$
$$\sum_{m=1}^{M}||\mathbf{w}_n^m||_2^2+||\mathbf{q}_n||_2^2\le P_{\text{TX},n}^{\max}, \quad \forall n\in\mathcal{N}. \tag{4.151}$$

By investigating (4.142)-(4.151), it can be observed that the objectives and most constraints are the same as the case when perfect CSI is available. The only difference lies at the QoS requirements (4.143) and (4.148). They guarantee that, as long as the CSI error is bounded, even in the worst case, the QoS targets can still be achieved for each UE, even though the error is not known exactly. It can also be interpreted in another way: For UE k, as the error is bounded within the sphere of radius ϵ_k with the probability of at least $1-\delta_k$, the QoS requirement (4.143) and (4.148) ensure that the QoS can be achieved, with the probability of at least $1-\delta_k$. By solving the problems above, a robust design of the network can be acquired. However, as the random and unknown $\{\mathbf{e}_k^{\text{CSI}}\}_{k=1}^K$ cannot be manipulated, getting rid of these parameters is necessary, in order to make the problems solvable. To avoid repetitions, we select the soft transfer mode with dedicated fronthaul as an example to elaborate on the algorithm, the extension to other cases can be followed by the way we have introduced before.

When the problem consisting of (4.147)-(4.149) and (4.151) is to be solved, the only difficulty lies in (4.148), as the others can be easily convexified with SDR and the iterative ℓ_0-norm approximation method, which have been introduced in previous sections. The key to deal with (4.148) is the adoption of the *S-Lemma*, which has been introduced in Subsection 2.3.6. For ease of explanation, we repeat the S-Lemma here.

S-Lemma: Let two functions $f_0(\mathbf{x}), f_1(\mathbf{x})$ defined as $f_0(\mathbf{x})=\mathbf{x}^H\mathbf{A}_0\mathbf{x}+2\text{Re}\{\mathbf{x}^H\mathbf{b}_0\}+c_0$ and $f_1(\mathbf{x})=\mathbf{x}^H\mathbf{A}_1\mathbf{x}+2\text{Re}\{\mathbf{x}^H\mathbf{b}_1\}+c_1$, where $\mathbf{b}_0,\mathbf{b}_1\in\mathbb{C}^{d\times 1}$ denote vectors, ma-

trices $\mathbf{A}_0, \mathbf{A}_1 \in \mathbb{C}^{d\times d}$ are all Hermitian matrices; and c_0, c_1 are scalars. Suppose that a specific vector $\hat{\mathbf{x}} \in \mathbb{C}^{d\times 1}$ exists, with which $f_1(\hat{\mathbf{x}}) < 0$ is satisfied. Then $f_0(\mathbf{x}) \geq 0$ and $f_1(\mathbf{x}) \leq 0$ can be satisfied simultaneously, for **arbitrary** $\mathbf{x} \in \mathbb{C}^{d\times 1}$, as long as a scalar $\lambda \geq 0$ exists, which makes the following matrix positive semi-definite, i.e.,

$$\begin{bmatrix} \mathbf{A}_0 & \mathbf{b}_0 \\ \mathbf{b}_0^H & c_0 \end{bmatrix} + \lambda \begin{bmatrix} \mathbf{A}_1 & \mathbf{b}_1 \\ \mathbf{b}_1^H & c_1 \end{bmatrix} \succeq \mathbf{0}. \tag{4.152}$$

At first, with the SDR technique, (4.148) can be reformulated into the following form where the S-Lemma can be applied:

$$\begin{aligned} &\frac{1}{\Gamma^m}\mathbf{h}_k^H\mathbf{W}^m\mathbf{h}_k - \sigma_k^2 - \max_{||\mathbf{e}_k^{\text{CSI}}||_2^2 \leq \epsilon_k^2} \Bigg(\left(\mathbf{h}_k^H + \mathbf{e}_k^{\text{CSI}^H}\right)\mathbf{Q}\left(\mathbf{h}_k + \mathbf{e}_k^{\text{CSI}}\right) + \mathbf{e}_k^{\text{CSI}^H}\mathbf{W}^m\mathbf{e}_k^{\text{CSI}} \\ &+ \sum_{i\neq m}^{M}\left(\mathbf{h}_k^H + \mathbf{e}_k^{\text{CSI}^H}\right)\mathbf{W}^i\left(\mathbf{h}_k + \mathbf{e}_k^{\text{CSI}}\right)\Bigg) \geq 0, \quad \forall k \in \mathcal{G}^m, \quad \forall m \in \mathcal{M}. \end{aligned} \tag{4.153}$$

Similarly, in the expression above, $\mathbf{W}^m = \mathbf{w}^m\mathbf{w}^{mH} \in \mathbb{R}^{NL\times NL}\ \forall m \in \mathcal{M}$ and $\mathbf{Q} = \mathbf{q}\mathbf{q}^H \in \mathbb{R}^{NL\times NL}$ are positive semidefinite matrices, i.e., $\mathbf{Q}$, $\{\mathbf{W}^m\}_{m=1}^M \succeq \mathbf{0}$. For any fixed $\mathbf{e}_k^{\text{CSI}}$, the LHS of (4.153) is a convex function with respect to $\mathbf{Q}$ and $\{\mathbf{W}^m\}_{m=1}^M$. Note that (4.153) is only a relaxed version of (4.148), as the non-convex constraints $\text{rank}\,(\mathbf{W}^m) = 1\ \forall m \in \mathcal{M}$ and $\text{rank}\,(\mathbf{Q}) = 1$ are temporarily dropped. Then we adopt the S-Lemma: By introducing scalar auxiliary variables $\{\alpha_k, \beta_k, \gamma_k\}$, the inequality (4.153) can be equivalently expressed as

$$\frac{1}{\Gamma^m}\mathbf{h}_k^H\mathbf{W}^m\mathbf{h}_k - \sigma_k^2 - \alpha_k - \beta_k - \gamma_k \geq 0, \tag{4.154}$$

$$-\left(\mathbf{h}_k^H + \mathbf{e}_k^{\text{CSI}^H}\right)\mathbf{Q}\left(\mathbf{h}_k + \mathbf{e}_k^{\text{CSI}}\right) + \alpha_k \geq 0, \tag{4.155}$$

$$-\mathbf{e}_k^{\text{CSI}^H}\mathbf{W}^m\mathbf{e}_k^{\text{CSI}} + \beta_k \geq 0, \tag{4.156}$$

$$-\left(\mathbf{h}_k^H + \mathbf{e}_k^{\text{CSI}^H}\right)\left(\sum_{i\neq m}^{M}\mathbf{W}^i\right)\left(\mathbf{h}_k + \mathbf{e}_k^{\text{CSI}}\right) + \gamma_k \geq 0, \tag{4.157}$$

$$\mathbf{e}_k^{\text{CSI}^H}\mathbf{e}_k^{\text{CSI}} - \epsilon_k^2 \leq 0, \quad \forall k \in \mathcal{G}^m, \quad \forall m \in \mathcal{M}. \tag{4.158}$$

Then by adopting the S-Lemma to (4.155)-(4.158) [17], the original constraint (4.153)

[17] For example, by regarding $\mathbf{e}_k^{\text{CSI}}$ as $\mathbf{x}$ in the S-Lemma, the LHS of (4.155) as $f_0(\mathbf{x})$ and the LHS of (4.158) as $f_1(\mathbf{x})$, (4.155) is equivalent to (4.160) and $\lambda_k \geq 0$.

can be further equivalently expressed with the following constraints:

$$\frac{1}{\Gamma^m}\mathbf{h}_k^H\mathbf{W}^m\mathbf{h}_k - \sigma_k^2 - \alpha_k - \beta_k - \gamma_k \geq 0, \tag{4.159}$$

$$\mathbf{D}_k = \begin{bmatrix} \lambda_k\mathbf{I}_{NL\times NL} - \mathbf{Q} & -\mathbf{Q}\mathbf{h}_k \\ -\mathbf{h}_k^H\mathbf{Q}^H & -\lambda_k\epsilon_k^2 + \alpha_k \end{bmatrix} \succeq \mathbf{0}, \tag{4.160}$$

$$\mathbf{E}_k = \begin{bmatrix} \mu_k\mathbf{I}_{NL\times NL} - \mathbf{W}^m & \mathbf{0}_{NL\times 1} \\ \mathbf{0}_{NL\times 1}^H & -\mu_k\epsilon_k^2 + \beta_k \end{bmatrix} \succeq \mathbf{0}, \tag{4.161}$$

$$\mathbf{F}_k = \begin{bmatrix} \lambda_k\mathbf{I}_{NL\times NL} - \sum_{i\neq m}^M \mathbf{W}^i & \left(-\sum_{i\neq m}^M \mathbf{W}^i\right)\mathbf{h}_k \\ \mathbf{h}_k^H\left(-\sum_{i\neq m}^M \mathbf{W}^{iH}\right) & -\nu_k\epsilon_k^2 - \mathbf{h}_k^H\left(\sum_{i\neq m}^M \mathbf{W}^i\right)\mathbf{h}_k + \gamma_k \end{bmatrix} \succeq \mathbf{0}, \tag{4.162}$$

$$\lambda_k,\ \mu_k,\ \nu_k \geq 0, \qquad \forall k \in \mathcal{G}^m, \quad \forall m \in \mathcal{M}. \tag{4.163}$$

Obviously, with constraints (4.159)-(4.163), we finally get rid of the unknown $\{\mathbf{e}_k^{\text{CSI}}\}_{k=1}^K$, but replace them with the known $\{\epsilon_k\}_{k=1}^K$. Moreover, these constraints are convex with respect to both the parameters to be optimized, and the introduced auxiliary parameters. Together with adopting the techniques introduced in previous sections, for the convexification of the objective (4.147) and the other constraints (4.149)-(4.151), the original problem can be equivalently reformulated as follows:

$$\mathcal{P}^{\text{Soft}}_{\text{Inaccurate CSI}}: \min_{\{\mathbf{W}^m\}_{m=1}^M, \mathbf{Q}} \sum_{m=1}^{M} \text{tr}(\mathbf{W}^m) + \sum_{n=1}^{N} \text{tr}(\mathbf{Q}\mathbf{J}_n) + \sum_{n=1}^{N} \Delta P \Big| \sum_{m=1}^{M} \text{tr}(\mathbf{W}^m \mathbf{J}_n) \Big|_0, \tag{4.164}$$

$$\text{s.t.} \quad \frac{1}{\Gamma^m} \mathbf{h}_k^H \mathbf{W}^m \mathbf{h}_k - \sigma_k^2 - \alpha_k - \beta_k - \gamma_k \geq 0, \tag{4.165}$$

$$\mathbf{D}_k = \begin{bmatrix} \lambda_k \mathbf{I}_{NL\times NL} - \mathbf{Q} & -\mathbf{Q}\mathbf{h}_k \\ -\mathbf{h}_k^H \mathbf{Q}^H & -\lambda_k \epsilon_k^2 + \alpha_k \end{bmatrix} \succeq \mathbf{0}, \tag{4.166}$$

$$\mathbf{E}_k = \begin{bmatrix} \mu_k \mathbf{I}_{NL\times NL} - \mathbf{W}^m & \mathbf{0}_{NL\times 1} \\ \mathbf{0}_{NL\times 1}^H & -\mu_k \epsilon_k^2 + \beta_k \end{bmatrix} \succeq \mathbf{0}, \tag{4.167}$$

$$\mathbf{F}_k = \begin{bmatrix} \lambda_k \mathbf{I}_{NL\times NL} - \sum_{i\neq m}^{M} \mathbf{W}^i & \left(-\sum_{i\neq m}^{M} \mathbf{W}^i\right) \mathbf{h}_k \\ \mathbf{h}_k^H \left(-\sum_{i\neq m}^{M} \mathbf{W}^{iH}\right) & -\nu_k \epsilon_k^2 - \mathbf{h}_k^H \left(\sum_{i\neq m}^{M} \mathbf{W}^i\right) \mathbf{h}_k + \gamma_k \end{bmatrix} \succeq \mathbf{0}, \tag{4.168}$$

$$\sum_{l=1}^{L} \left(\log_2 \eta_{n,l} + \frac{\text{tr}\left(\mathbf{Q}^{(t+1)} \mathbf{J}_{n,l}\right) + \sum_{m=1}^{M} (1 - c_n^{f^m}) \text{tr}\left(\mathbf{W}^{m(t+1)} \mathbf{J}_{n,l}\right)}{\eta_{n,l} \ln 2} \right)$$
$$- \sum_{l=1}^{L} \log_2 \text{tr}\left(\mathbf{Q}^{(t+1)} \mathbf{J}_{n,l}\right) - \frac{L}{\ln 2} - C_{\text{FH},n} \leq 0, \quad \forall n \in \mathcal{N}, \tag{4.169}$$

$$\sum_{m=1}^{M} \text{tr}(\mathbf{W}^m \mathbf{J}_n) + \text{tr}(\mathbf{Q}\mathbf{J}_n) \leq P_{\text{TX},n}^{\max}, \quad \forall n \in \mathcal{N}, \tag{4.170}$$

$$\mathbf{W}^m \succeq \mathbf{0}, \quad \forall m \in \mathcal{M}, \tag{4.171}$$

$$\mathbf{Q} \succeq \mathbf{0}, \tag{4.172}$$

$$\text{rank}\,(\mathbf{W}^m) = 1, \quad \forall m \in \mathcal{M} \tag{4.173}$$

$$\text{rank}\,(\mathbf{Q}) = 1, \tag{4.174}$$

$$\lambda_k,\ \mu_k,\ \nu_k \geq 0, \qquad \forall k \in \mathcal{G}^m, \quad \forall m \in \mathcal{M}, \tag{4.175}$$

where ξ in (4.164) denotes the power amplifier efficiency and $\Delta P = \xi(P_{\text{o}} - P_{\text{sleep}})$. The constraints (4.169) are obtained with the same upper-bounding technique used in Subsection 4.2.1.2, for details please refer to (4.51)-(4.60).

After dropping the rank constraints (4.173) and (4.174), the problem above is again a standard SDP problem, whose solving procedure is in line with Alg. 3. Similarly, an initial SDP problem has to be constructed and solved, which shall be expressed

as follows:

$$\mathcal{P}^{\text{Soft }(0)}_{\text{Inaccurate CSI}}: \min_{\{\mathbf{W}^{m(0)}\}_{m=1}^{M},\mathbf{Q}^{(0)}} \sum_{m=1}^{M} \text{tr}(\mathbf{W}^{m(0)}) + \sum_{n=1}^{N} \text{tr}(\mathbf{Q}^{(0)}\mathbf{J}_n), \tag{4.176}$$

$$\text{s.t.} \quad \frac{1}{\Gamma^m}\mathbf{h}_k^H\mathbf{W}^{m(0)}\mathbf{h}_k - \sigma_k^2 - \alpha_k^{(0)} - \beta_k^{(0)} - \gamma_k^{(0)} \geq 0, \tag{4.177}$$

$$\mathbf{D}_k^{(0)} = \begin{bmatrix} \lambda_k^{(0)}\mathbf{I}_{NL\times NL} - \mathbf{Q}^{(0)} & -\mathbf{Q}^{(0)}\mathbf{h}_k \\ -\mathbf{h}_k^H\mathbf{Q}^{(0)H} & -\lambda_k^{(0)}\epsilon_k^2 + \alpha_k^{(0)} \end{bmatrix} \succeq \mathbf{0}, \tag{4.178}$$

$$\mathbf{E}_k^{(0)} = \begin{bmatrix} \mu_k^{(0)}\mathbf{I}_{NL\times NL} - \mathbf{W}^{m(0)} & \mathbf{0}_{NL\times 1} \\ \mathbf{0}_{NL\times 1}^H & -\mu_k^{(0)}\epsilon_k^2 + \beta_k^{(0)} \end{bmatrix} \succeq \mathbf{0}, \tag{4.179}$$

$$\mathbf{F}_k^{(0)} = \begin{bmatrix} \lambda_k^{(0)}\mathbf{I}_{NL\times NL} - \sum_{i\neq m}^{M}\mathbf{W}^{i(0)} & \left(-\sum_{i\neq m}^{M}\mathbf{W}^{i(0)}\right)\mathbf{h}_k \\ \mathbf{h}_k^H\left(-\sum_{i\neq m}^{M}\mathbf{W}^{i(0)H}\right) & -\nu_k^{(0)}\epsilon_k^2 - \mathbf{h}_k^H\left(\sum_{i\neq m}^{M}\mathbf{W}^{i(0)}\right)\mathbf{h}_k + \gamma_k^{(0)} \end{bmatrix} \succeq \mathbf{0}, \tag{4.180}$$

$$\sum_{m=1}^{M} \text{tr}(\mathbf{W}^{m(0)}\mathbf{J}_n) + \text{tr}(\mathbf{Q}^{(0)}\mathbf{J}_n) \leq P_{\text{TX},n}^{\max}, \quad \forall n \in \mathcal{N}, \tag{4.181}$$

$$\mathbf{W}^{m(0)} \succeq \mathbf{0}, \quad \forall m \in \mathcal{M}, \tag{4.182}$$

$$\mathbf{Q}^{(0)} \succeq \mathbf{0}, \tag{4.183}$$

$$\lambda_k^{(0)},\ \mu_k^{(0)},\ \nu_k^{(0)} \geq 0, \qquad \forall k \in \mathcal{G}^m, \quad \forall m \in \mathcal{M}, \tag{4.184}$$

And for the $(t+1)$-iteration afterwards, the problem $\mathcal{P}^{\text{Soft }(t+1)}_{\text{Inaccurate CSI}}$ to be solved is constructed according to (4.164)-(4.175), but without (4.173) and (4.174).

In summary, the robust design procedure for the soft transfer mode is documented in Alg. 7.

4.3.2 High SE oriented Robust Design

By reviewing the two algorithms proposed for the high SE oriented design with perfect CSI, Alg. 5 and Alg. 6, we see that both of them rely on solving a related power minimization problem. Obviously, Alg. 5 is rather easy to be extended to the case with inaccurate CSI, as it is only a combination of the Bi-Section method, and the solution for the power minimization problem. Hence, when only inaccurate CSI is available, we can just replace the constructed power minimization problem in Alg. 5, with the power minimization problem with inaccurate CSI. The algorithm for solving such a problem has been provided in the last subsection, where we took the soft transfer mode as an example. However, when the multi-cast throughput maximization is considered, two sub-problems are constructed and solved: The first one is in the Re-Design sub-step, which is also a power minimization problem. However, in the Re-Allocation sub-step, the power allocation algorithm cannot be easily

Algorithm 7: The Iterative Optimization Steps for Robust TX power Minimization (*For the soft transfer mode*)

Initialization: Solve the standard SDP problem $\mathcal{P}^{\text{Soft }(0)}_{\text{Inaccurate CSI}}$ (4.176)-(4.184) to obtain $\{\mathbf{W}^{m(0)}\}_{m=1}^{M}$ and $\mathbf{Q}^{(0)}$. Compute $\eta_{n,l}^{(1)}$ based on (4.60), $\forall n, l$. Construct the problem $\mathcal{P}^{\text{Soft }(1)}_{\text{Inaccurate CSI}}$ according to (4.164)-(4.175) without (4.173) and (4.174), and set $t \leftarrow 1$.
repeat
  Solve the standard SDP problem $\mathcal{P}^{\text{Soft }(t)}_{\text{Inaccurate CSI}}$ for obtaining $\{\mathbf{V}^{m(t)}\}_{m=1}^{M}$ and $\mathbf{Q}^{(t)}$.
  Compute the values of $\eta_{n,l}^{(t+1)}$ based on (4.60), $\forall n, l$. Then formulate the problem $\mathcal{P}^{\text{Soft }(t+1)}_{\text{Inaccurate CSI}}$ according to (4.164)-(4.175) without (4.173) and (4.174), and set $t \leftarrow t+1$.
until *convergence or reaching the max iteration number*;
if $\text{rank}(\mathbf{W}^{m(\text{last})}) = 1$ *and* $\text{rank}(\mathbf{Q}^{(\text{last})}) = 1$ **then**
  Perform EVD to obtain the optimal $\{\mathbf{w}^{m}\}_{m=1}^{M}$ and $\mathbf{q}$.
else
  Use Gaussian randomization and scaling [KSL08] method to obtain the approximate solution $\{\mathbf{w}^{m}\}_{m=1}^{M}$ and $\mathbf{q}$.

adapted with inaccurate CSI, as in the sub-gradient method, exact values of SINR need to be computed, as shown in (4.137) and (4.138). With unknown $\{\mathbf{e}_k^{\text{CSI}}\}_{k=1}^{K}$, it is impossible to compute the value, as shown in (4.140) and (4.141). Therefore, Alg. 6 is not possible to be used for scenarios with inaccurate CSI. Fortunately, as shown in Subsection 4.2.5, when the network resources are abundant, i.e., with large cache memory size S or fronthaul capacity C_{FH}, the results of wMMF are close to TP-Max. So the low-complexity Alg. 5 can be adopted to approach the results for the multicast throughput maximization, in scenarios where CSI is inaccurate and network resources are abundant.

4.3.3 Numerical Results

In this subsection the numerical results for the robust design are to be provided. Again, the same simulation parameters and methods are adopted as before. In simulations, the hard transfer mode is adopted and the same distortion level is assumed for all UEs, i.e., $\epsilon = \epsilon_k \ \forall k$. Moreover, we select $\delta_k = 0 \ \forall k$ for easier illustration, i.e., the distortion is assumed to be always bounded without outage probability. For the case when $\delta_k > 0$, the proposed algorithm will run into outage with the probability of δ, but all conclusions below are still valid and the algorithm keeps the same.

At first we show how such a robust design influences the network power consumption: Among all independent realizations, we randomly pick up one, and compare the recorded transmission power of each eRRH for each iteration of our algorithm.

Figure 4.33: eRRH deactivation for perfect CSI.

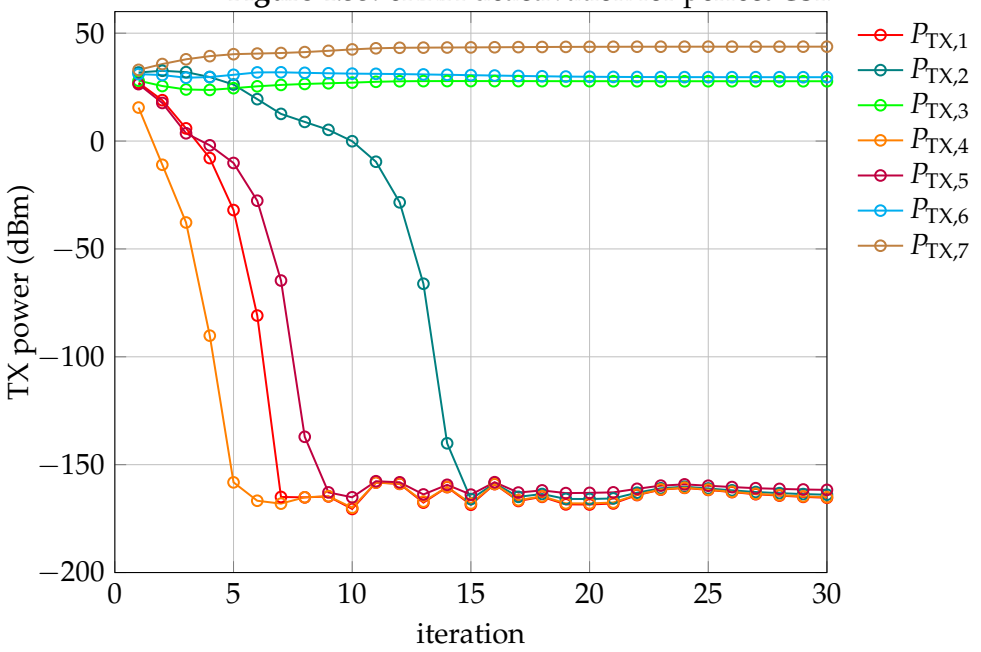
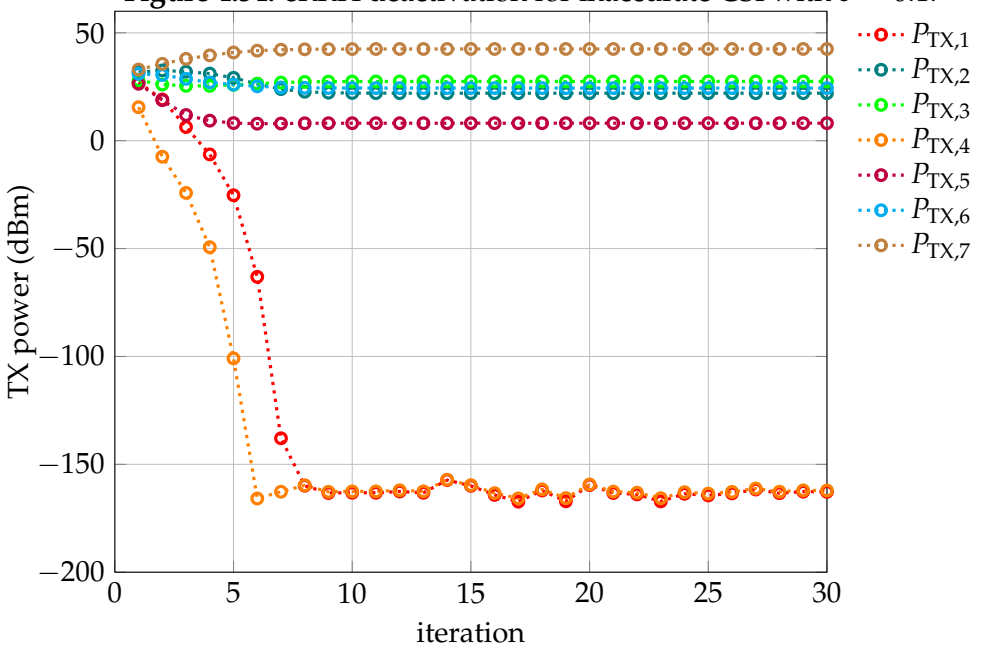


Figure 4.34: eRRH deactivation for inaccurate CSI with $\epsilon = 0.1$.



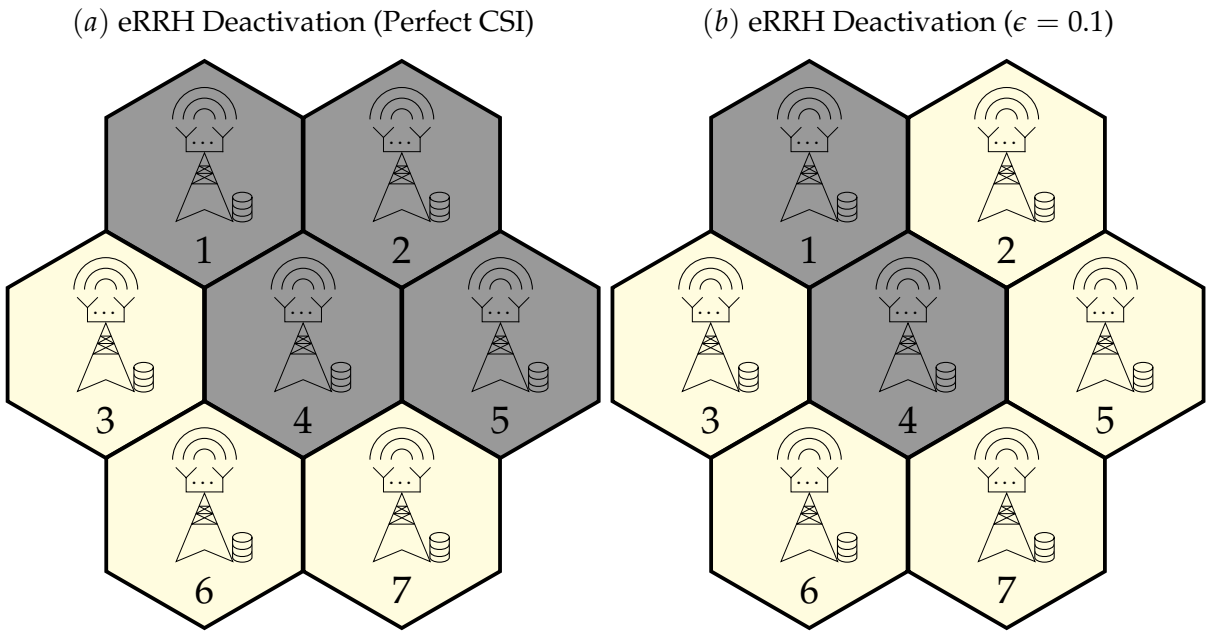


Figure 4.35: An illustration of the final eRRH deactivation results for the case when perfect CSI is available and CSI is distorted with $\epsilon = 0.1$. Cell colored with gray denotes that the eRRH within this cell is deactivated.

Remember that when the value of the transmission power of a specific eRRH drops below the threshold parameter τ (in our case it is -50 dBm), this eRRH is determined to be deactivated, more details have been stated in Subsection 4.2.2. In Fig. 4.33 and Fig. 4.34, the results are compared for the case of perfect CSI and the case of inaccurate CSI with distortion level $\epsilon = 0.1$, which is a relatively large value when measuring the distortion.

It can be observed that with the proposed algorithm, all eRRHs are active at the beginning, but some of them are deactivated gradually for saving power. From the figure, we see that four eRRHs (eRRH 1, eRRH 2, eRRH 4 and eRRH 5) are switched off in 15 iterations, as the corresponding power falls far below the threshold -50 dBm, when perfect CSI is available. However, when the robust design is executed with inaccurate CSI knowledge, only two eRRHs (eRRH 1 and eRRH 4) can be deactivated. This is due to more network resources (incl. power, fronthaul capacities, caches, etc.) are required, in order to counteract the network uncertainties to guarantee the robustness. Based on the results above, a more intuitive comparison between this these cases is illustrated in Fig. 4.35.

Then the overall performance is investigated instead of a specific slot realization: We set up 200 independent realizations and execute the algorithms for both cases of perfect and inaccurate CSI, then the number of eRRHs that are still active after the algorithms terminate, i.e., after 20 iterations, are documented. After averaging these numbers, the probability distribution of the number of active eRRHs is computed and depicted in Fig. 4.36. Obviously, when perfect CSI is available, more eRRHs

have better chance to be deactivated, which is in line with the results shown in Fig. 4.33 for that specific realization. For example, when the cache memory size $S = 0$ and perfect CSI is available, the network has nearly 40% probability to turn off two eRRHs, such that the remaining five can still fulfill the QoS targets of all UEs. There is only a probability of 20%, such that all eRRHs must be activated, which means that some eRRHs (at least one) can be deactivated to save power with the probability of 80%. However, when only inaccurate CSI is available, such a probability decreases to about 20%. In most cases ($\sim$ 80%), all seven eRRHs have to keep active. Furthermore, we can also observe that when there are more available network resources, e.g., larger cache memory size S, more eRRHs have the possibility to be switched off, and it holds for both perfect and inaccurate CSI scenarios.

Next, we verify whether with the proposed algorithm, the network is indeed robust, such that even perfect CSI is not known, the QoS at each UE can still be guaranteed. Firstly, a new metric called the *normalized rate* is set up :

$$R_k^{\text{Norm}} = \frac{\log_2(1+\text{SINR}_k)}{\log_2(1+\Gamma^m)}, \tag{4.185}$$

in which SINR_k is the actually achieved SINR of UE k, which is calculated according to (4.15) and (4.16) for the hard and soft transfer mode. Note that such a value is not available at the BBU pool as exact CSI is not known (but each UE can measure it), the robust beamformers/precoders are optimized by the BBU pool with inaccurate CSI. By substituting the resultant robust beamformers/precoders from the proposed algorithm, and the actual channel vectors into (4.15) and (4.16), the actual SINRs that are achieved at each UE can be computed. The normalized rate of UE k, i.e., R_k^{Norm}, is the ratio of the actually achieved rate $\log_2(1+\text{SINR}_k)$ to the QoS target $\log_2(1+\Gamma^m)$. If $R_k^{\text{Norm}} \geq 1 \; \forall n \in \mathcal{N}$ is satisfied, it can be claimed that the proposed algorithm indeed ensures the robustness, as the QoS of each UE is guaranteed. The normalized rate for each independent realization for different channel distortion levels are documented, and the probability distributions of them are illustrated in Fig. 4.37. As a comparison, the results of the non-robust algorithms are also provided. For such a non-robust design, the BBU pool just regards the distorted CSI as the exact one and optimize the network accordingly, with the algorithms introduced in Subsection 4.2.2. From the results depicted in Fig. 4.37, it is obvious that the robust algorithm always guarantees the QoS of each UE, as the normalized rates are 100% equal or larger than 1. These values are often larger than 1, since the robust design guarantees the worst case scenario: As long as the distortion is bounded, the QoS can be satisfied. By increasing the distortion level ϵ, the distribution becomes more spread in x-axis, as the uncertainty of the CSI knowledge is increased. However, the price to counteract more uncertainty is more power con-

Figure 4.36: The probability distribution of the averaged number of active eRRHs, with different cache memory sizes and channel distortion levels.

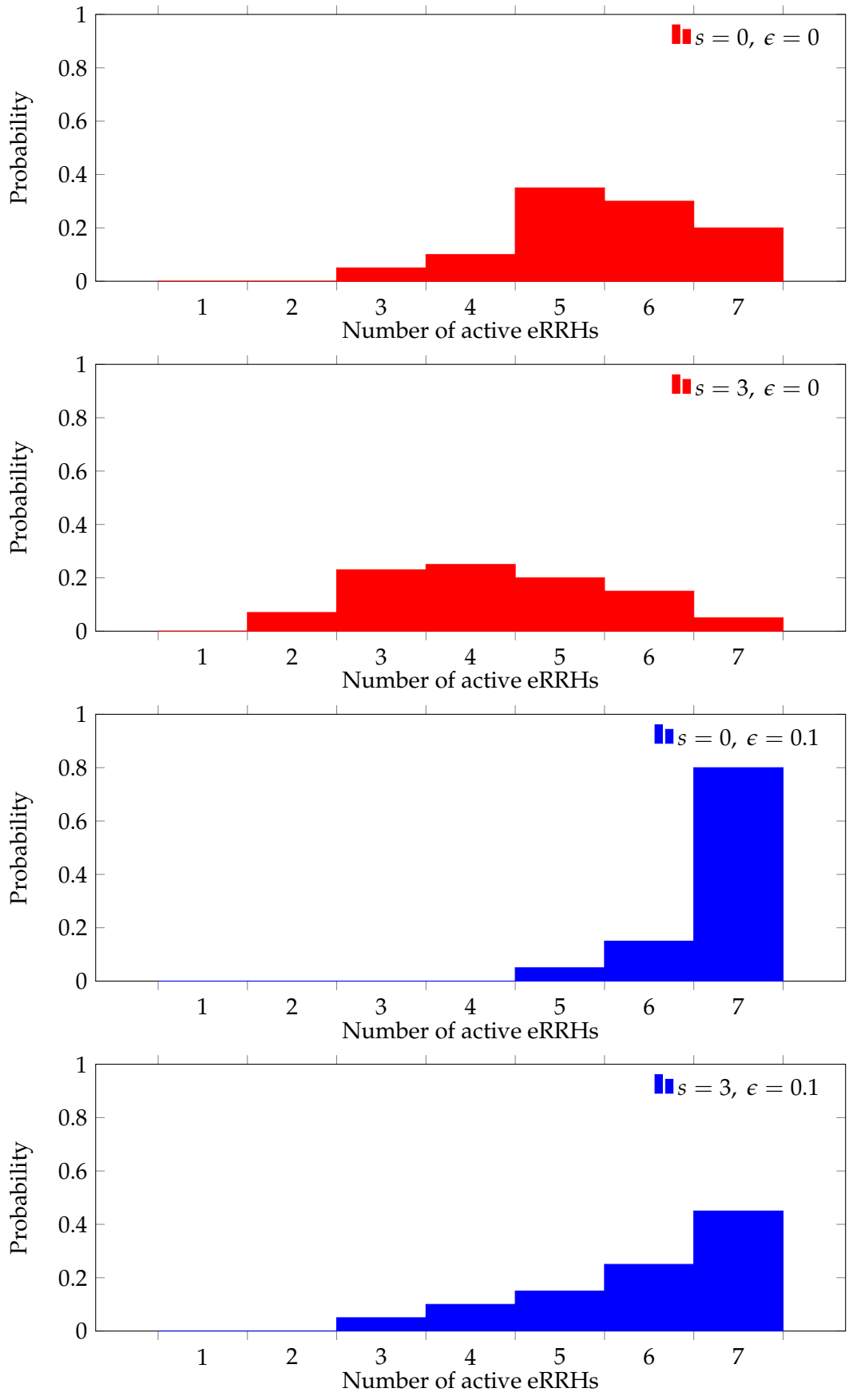


Figure 4.37: The probability distribution of the normalized rate for the robust and the non-robust design with different CSI distortion levels (QoS target Γ = 5 dB).

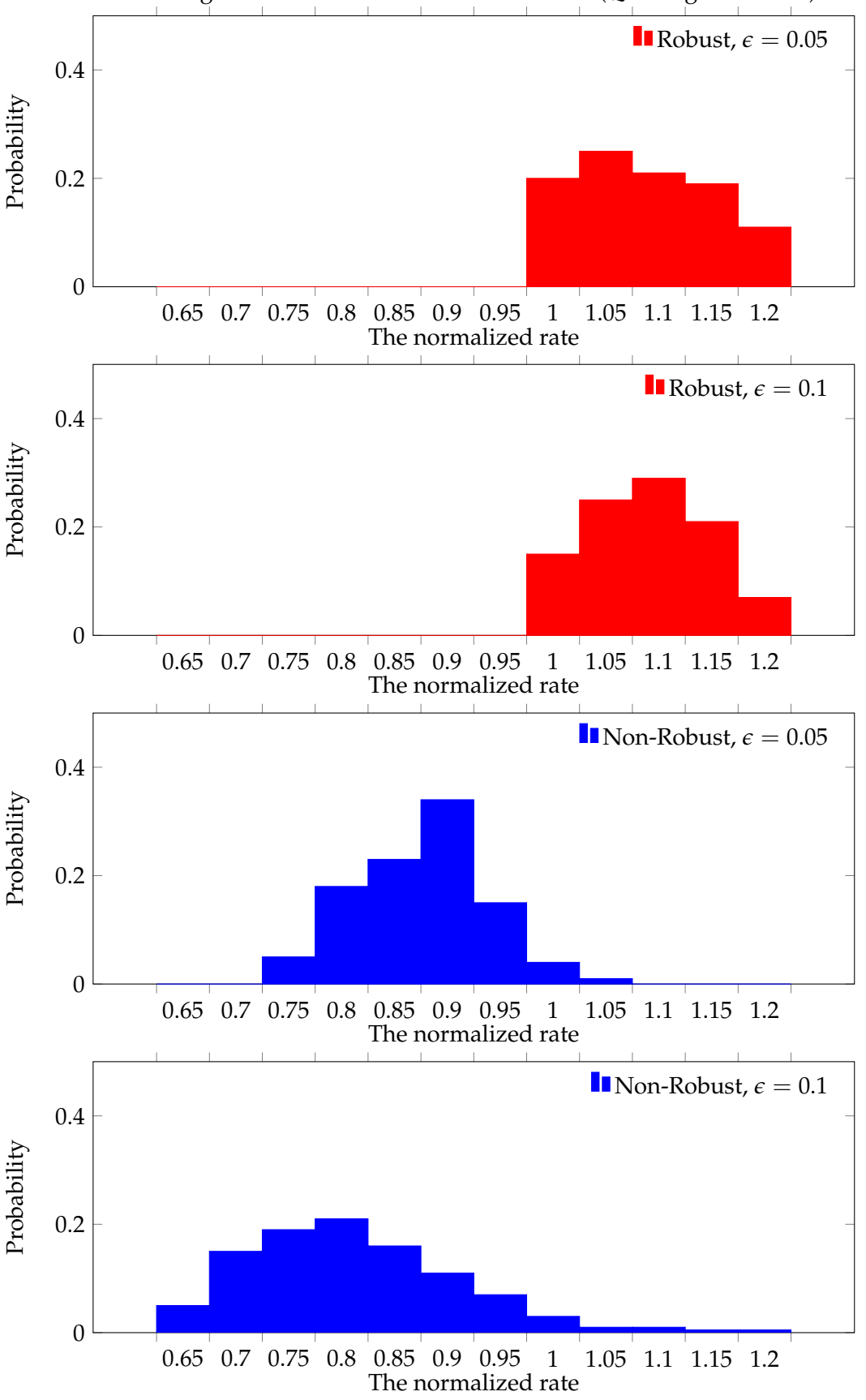
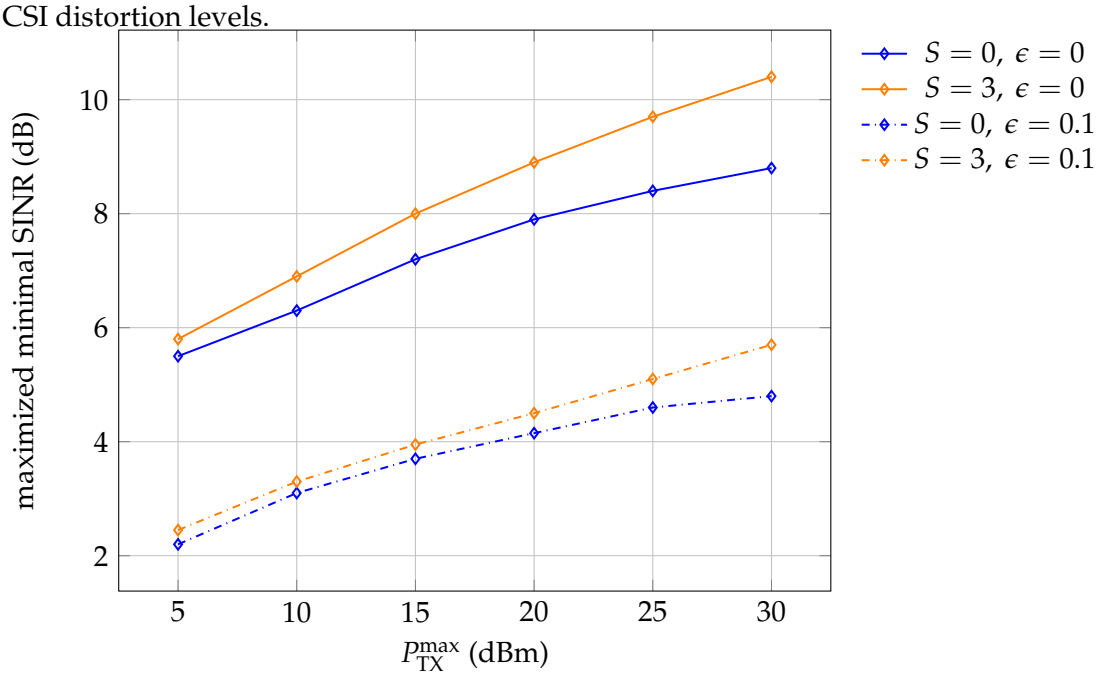


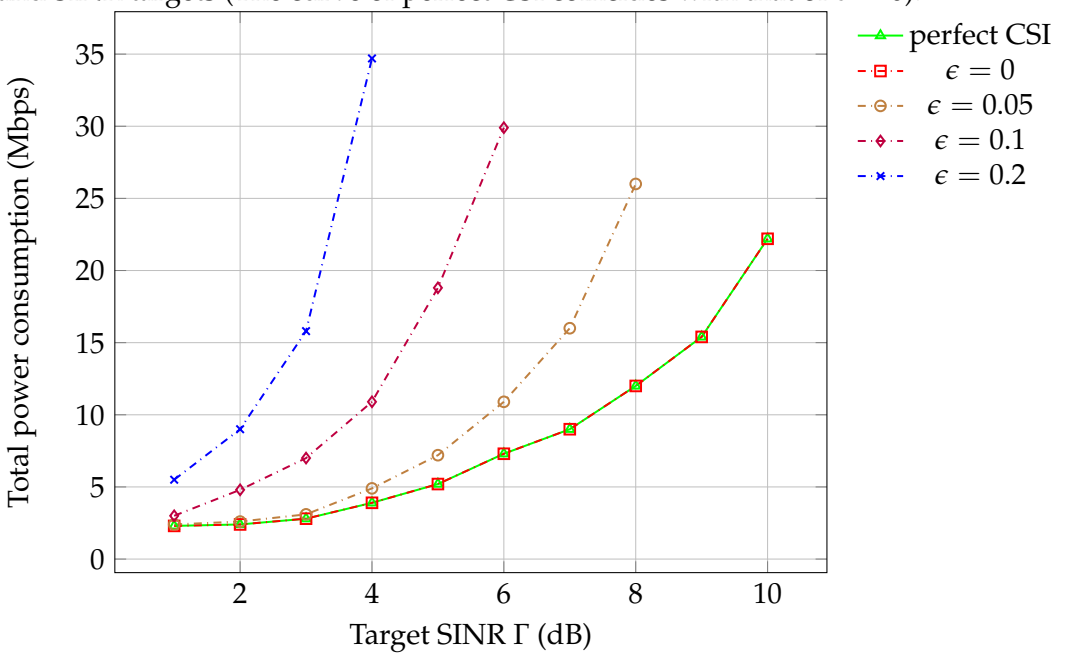
Figure 4.38: The maximized minimal SINR for different network configurations and CSI distortion levels.



sumption, which will be shown later. When it goes to the non-robust design, i.e., the BBU pool optimizes the network based on inaccurate CSI but regards them as accurate, the QoS of each UE cannot be guaranteed: When $\epsilon = 0.05$, only about 5% UEs among all, can get its desired QoS fulfilled. When $\epsilon = 0.1$, such a value decreases to around 3%. The results demonstrate the significance of the proposed robust design approach when only inaccurate CSI is available.

Concerning the robust high SE oriented design, Fig. 4.38 illustrates the maximized minimal SINR among all UEs, by averaging the results from all independent realization. We see that with higher maximal allowable power or larger cache memory size, higher achievable rates can be guaranteed for the UE with the worst channel conditions. The benefit of increasing the cache memory size becomes more and more manifest, as more power becomes available. This is because when less transmission power is available, the performance is limited by the radio access hop from the eRRHs to UEs, instead of the fronthaul. In such cases, the fronthaul resources are abundant, such that fronthauling contents to more eRRHs, in order to increase the array gain, is more probable, thus the benefit of caching contents is less significant. When more power is available, the performance is more and more limited by the fronthaul, which makes caching more beneficial. When only inaccurate CSI is available, more power is required at each eRRH to combat against the channel uncertainty, the network becomes more probable, to be limited by the power, instead of the fronthaul. Hence, in the case of inaccurate CSI, for a given power budget, the performance gap between whether cache exists, is smaller than the case with perfect CSI. While we can still conclude that, introducing cache module is a cheap

Figure 4.39: The minimized power consumption for different CSI distortion levels and SINR targets (The curve of perfect CSI coincides with that of $\epsilon = 0$).



solution with low-complexity, to increase the robustness of the network.

Fig. 4.39 illustrates the price that the network pays for such a robust design: Higher power consumption. When CSI becomes less accurate, i.e., the value of ϵ gets larger, much more transmission power is required to counteract the uncertainty, when the network aims to guarantee the same target SINR at the UE side. Sometimes the problem becomes even infeasible: With the current network configuration, it is not possible to robustly guarantee the QoS at each UE. For example, when $\epsilon = 0.2$, the problem becomes infeasible if the target SINR Γ larger than 4 dB, meaning that the current maximal allowable power and the fronthaul capacity cannot robustly support the QoS of each UE anymore, when $\Gamma > 4$ dB. Moreover, it should be noted that the results of *perfect CSI* are compared with that of $\epsilon = 0$: For perfect CSI, the algorithms proposed in Subsection 4.2.2 is executed, while for $\epsilon = 0$, the robust algorithm proposed in Subsection 4.3.1 is executed, but just by setting $\epsilon = 0$. The results of them, as expected, coincide with each other.

4.4 Discussions, Summaries, and Outlooks

In this chapter, the optimal network design for the downlink of F-RAN is investigated: Both high Energy Efficiency (EE) and Spectral Efficiency (SE) oriented design are discussed. Moreover, the robust design scheme is also studied, when only inaccurate CSI is available.

We start with the simplest case: Transmission power minimization for high EE oriented design with perfect CSI, in which both the hard and the soft transfer mode on fronthauls, as well as both dedicated and non-dedicated fronthauls, are introduced and studied. The main technique to tackle this problem is the SemiDefinite Relaxation (SDR) method and the iterative ℓ_0-norm approximation scheme, with which the problem can be convexified into a standard SDP problem, which can be efficiently solved by many existing solvers. Then we extend the case to the minimization of the total power of the network: Not only the transmission power is taken into account, but also all other operational power. In this case, we show that it is possible to switch off some eRRHs for saving the overall power of the network, instead of activating them all to decrease only the transmission power. With the proposed algorithm, the eRRH deactivation can also be dynamically optimized by the BBU pool.

Then the high SE oriented design is considered. Two different design metrics are investigated: The weighted Max-Min Fairness (wMMF) and the multi-cast Throughput Maximization (TP-Max). We propose algorithms for both of them, and each one relies on the previously discussed power minimization problem. When wMMF is the target, it is shown that by combining the solving of the power minimization problem and the Bi-Section method, the corresponding problem can be solved in a tortuous manner. For the TP-Max, as both beamformers/precoders and the power allocation are to be optimized, an alternating mechanism is proposed, such that one variable type is alternatively fixed, and the other is to be optimized. When the power allocation scheme is optimized, the sub-gradient method is adopted. Both theoretic analysis and numerical results are given, in order to show the convergence behaviour of such alternating steps.

Furthermore, the performance of the hard and the soft transfer mode are also compared. The results demonstrate that the soft transfer mode has better capability to exploit the networks resources in most cases. When the network resources get more limited, the benefit of the soft transfer mode becomes more apparent. However, the price of it is its higher implementation and optimization complexity.

At last, we address the robust design for both high EE and SE oriented design, i.e., when only inaccurate CSI is available at the BBU pool, the network design that can still guarantee the QoS of each UE. The S-Lemma is adopted to deal with this problem. With the S-Lemma, the original problem can also be converted into a SDP problem, which is then solved with the techniques adopted in earlier sections.

Apart from the algorithms, the benefits of introducing the cache module at eRRHs are also demonstrated by many numerical results. With caches, the aggregated networked array gain can be increased for achieving higher spatial diversity, which

has the same effect as if higher allowable power budgets, or larger fronthaul capacities, are available. Note that the increase of the power budgets and the fronthaul capacities are usually rather expensive. Hence, with the low-cost and flexible cache module, both EE and SE of the network can be easily improved. Moreover, it also proved to have the capability to increase the robustness of the network, when only inaccurate CSI is available.

However, there is still some issues that need further research. For example, as described in Subsection 4.3.2, the proposed algorithms for the robust design cannot be extended to maximizing the network multi-cast throughput. What we have done there is to use the robust algorithm of the wMMF to approximate it when the network resources are abundant, i.e., when large cache memory size S or fronthaul capacity C_{FH} is available. Hence, for the TP-Max, an efficient algorithm is required for the robust design. Moreover, all the proposed algorithms have to be executed in an iterative manner. Although most of them can converge within ten iterations, it might be still not feasible for some real-time applications that require extremely low latency. Hence, algorithms with less computational requirements are worth to be investigated in future.

Chapter 5

Partially Decentralized Design with Partial CSI

This chapter contains

[1] In the last two Chapters, we have completely characterized the optimal design for both uplink and downlink of the cache-enabled F-RAN. In this chapter, we move one step further. Remember that in all proposed algorithms in last chapters, the BBU pool in the cloud requires the knowledge of the global CSI, so as to perform the centralized optimization. When the network is reciprocal, the BBU pool can use the same global CSI knowledge for both uplink and downlink. For nonreciprocal channels, the uplink CSI knowledge is acquired via Channel Sounding. Each UE must send the Sounding Reference Signal (SRS) [3GP18] via eRRHs to the BBU pool, with which the BBU pool can estimate the global channel quality. For the downlink, the BBU pool sends the CSI-RS signal for UEs to estimate the channel quality. Then all UEs have to feedback the estimated results via PUCCH to the BBU pool. Obviously, the estimation of the global CSI requires lots of overhead used for the reference signals and the feedback. Besides the huge amount of the overhead, the overall latency introduced by theses schemes is also a critical problem, especially for some real-time applications. Moreover, the centralized optimization procedures at the BBU pool might put a high computational burden on it. When more and more UEs are to be scheduled in each slot, the complexity might become unacceptable. Although such a burden can be relieved by introducing the fog computing, with which eRRHs can execute storage and computation tasks (e.g., the compression task

[1] Parts of this chapter have been published in [Che18].

in the uplink, or the recovery of the compressed signals in the downlink when the soft transfer mode is adopted), a centralized optimization is still needed. The complexity of such a centralized optimization grows exponentially with the number of eRRHs, as well as the number of antennas equipped on each eRRH. Hence, when more eRRHs are deployed in F-RAN to increase the coverage of the network, or equipping eRRHs with more antennas to further improve the performance, the actual performance of the network might not be as expected, since the complexity can exceed the capability of the BBU pool, and the increased overhead can overwhelm the benefit they bring. Hence, such drawbacks limit the network capability for more UE's coverage and performance improvement.

The purpose of this chapter is to overcome some of the drawbacks listed above. We will give some first trials by developing a partially decentralized algorithm with only partial CSI knowledge. Several parts of the computation tasks are carried by eRRHs via their fog computational capabilities, and based on only its local CSI knowledge. Hence, as we are going to show, the amount of overhead, as well as the computational burden at the BBU pool, can be greatly reduced. In particular, the computational complexity does not depend on the number of eRRHs within the network, as well as the number of antennas per eRRH. Instead, the complexity of the mechanism going to be proposed in this chapter, depends only on the number of UEs to be scheduled.

We emphasize here that the contents in this chapter cannot cover all topics, that have been discussed for the centralized approach in previous chapters. Only some first ideas will be presented and analyzed, so as to shed some lights on how to overcome some difficulties of the centralized design. More intensive work on this topic requires more research in future.

5.1 Introduction and System Model

5.1.1 Introduction

The key technology to achieve the partially decentralized algorithm is the concept of Massive MIMO, as have been introduced in Section 1.4. With Massive MIMO, a Base Station (BS) is equipped with a large number of antennas (e.g., 128 or 256). Such a technique relies on the law of large numbers: The large number of antennas can eliminate the effects of the small-scale fading and frequency dependence [Mar10; Mar+16]. From the perspective of an UE, the channel is hardened (channel hardening effect) to be a deterministic scalar channel, with known channel gain and additive noise. In another word, the exact CSI knowledge from the UE to each

antenna of the BS is not necessary anymore, when the achievable rate for this UE is considered. As a comparison, for all algorithms proposed in last chapters, such knowledge is necessary when the optimization is implemented. Then a natural question is: If such a property can be somehow adopted into the F-RAN, is it possible to reduce the requirements of delivering global CSI?

However, one significant issue of Massive MIMO is that its expected performance degrades rather rapid when the number of antennas is decreased. Hence, a single BS needs to be equipped with a large number of antennas in order to ensure the effectiveness of Massive MIMO. Such a requirement is hardly to be met for a micro BS. As much higher frequency bands are used by the 5G network, the 5G BS should be much more densely distributed to decrease the distances. Therefore, a 5G BS tends to be small enough for an easy and dense deployment, which might contradict with the requirements to achieve the desired performance of Massive MIMO. On the other hand, a F-RAN, consisting of a BBU pool and multiple eRRHs connected via fronthauls, forms a networked MIMO system in the charge of the cloud server. Now a natural question arises: Is it possible to achieve a networked Massive MIMO system with the help of the BBU pool, and multiple eRRHs? If yes, each eRRH might not need to be equipped with too many antennas, and some benefits of Massive MIMO can still be preserved.

Obviously, if these two techniques can be combined, i.e., F-RAN and Massive MIMO, they can potentially benefit from each other and overcome the drawbacks and limitations of themselves. We name such a combination **Networked Massive MIMO based F-RAN**, which is shown in Fig. 1.8. Similar to the F-RAN, it consists of a BBU pool in the cloud and multiple eRRHs at the network edge. They communicate with each other via fronthauls. However, each eRRH here is equipped with more antennas, such that the whole network can be regarded as a Massive MIMO system. But compared with a single Massive MIMO Base Station, each eRRH in this architecture does not need to be equipped with so many antennas. Then from the perspective of the BBU pool, some properties of the Massive MIMO can still be kept. In summary, such a combination has the following advantages, which we are going to elaborate next in detail:

1. It can reduce the amount of the data streams delivered by fronthauls, which scales with the number of the scheduled UEs, i.e., K, instead of the number of antennas L, and the number of eRRHs N, improving the performance with more antennas or eRRHs will not put much more burdens on the network;

2. The global instantaneous CSI knowledge is not required anymore at the BBU pool. Therefore, the amount of overhead exchanged within the network can be greatly reduced, especially when the number of antennas L, or the number of eRRHs N is large;

3. Compared with the centralized mechanisms introduced in last chapters, the proposed decentralized signal processing mechanism at eRRHs with its fog computing capability, can greatly reduce the complexity of the optimization, as well as the triggered latency;

4. The hardware costs at eRRHs can also be reduced as less compressors are needed.

5.1.2 System Model

We consider the uplink of the Massive MIMO based F-RAN, as depicted in Fig. 1.8. Totally K single-antenna UEs are scheduled to upload their contents to the BBU pool in the cloud, via N eRRHs and fronthauls. Each eRRH in $\mathcal{N} = \{1, 2, ..., N\}$ has limited signal processing capabilities, with which some distributed fog computing tasks can be executed. Each eRRH is equipped with a moderate numbers of antennas (e.g., 32), which is denoted by L. Such a value needs not to be so large compared with a typical Massive MIMO . Similarly, eRRH n connects to the BBU pool in the cloud with the fronthaul of capacity $C_{\text{FH},n}$.

Let ρ_{ul} be the maximal allowable uplink transmission power among all UEs, and s_k denote a realization of the transmitted symbol from UE $k \in \mathcal{K} = \{1, 2, ..., K\}$ with normalized power, $\eta_k \in [0, 1]$ denote the power control factor for UE k, i.e., how much power are used for UE k for the uplink transmission. Then the transmitted signal x_k from UE k, and the aggregated transmitted vector $\mathbf{x} \in \mathbb{C}^{K\times 1}$ among all UEs can be expressed as follows:

$$\begin{aligned} x_k &= \sqrt{\rho_{\text{ul}}\eta_k} s_k, \\ \mathbf{x} &= \sqrt{\rho_{\text{ul}}} \mathbf{D}_{\boldsymbol{\eta}}^{1/2} \mathbf{s}, \end{aligned} \tag{5.1}$$

where $\mathbf{D}_{\boldsymbol{\eta}} = \text{Diag}\left([\eta_1, \eta_2, ..., \eta_K]^T\right)$, $\mathbf{s} = [s_1, s_2, ..., s_K]^T$.

Let the channel gain from UE k to l-th antenna of eRRH n be $g_{n,k}^l$. According to [Mar+16], it can be further expressed as

$$g_{n,k}^l = h_{n,k}^l \sqrt{\beta_{n,k}}, \tag{5.2}$$

which consists of a large scale fading coefficient $\beta_{n,k}$ and a small scale fading coefficient $h_{n,k}^l$. Coefficient $\beta_{n,k}$ is determined by the distance between eRRH n and UE k (path loss), shadowing, etc. It varies relatively slow compared with the other coefficient, and it can be regarded the same between all antennas of eRRH n and UE k [Mar+16], as the distance between antennas of an eRRH is negligible compared with

the distance between an UE and an eRRH. In contrast, the small scale fading coefficient $h^l_{n,k}$ varies much faster and is independent among all antennas of an eRRH. Moreover, $h^l_{n,k}$ $\forall n,k,l$ are usually supposed to be Rayleigh distributed, i.e., i.i.d. $\mathcal{CN}(0,1)$ random variables, so we also adopt such assumptions here. The channel gains $g^l_{n,k}$ $\forall n,k,l$ are estimated at each antenna of each eRRH via the Sounding Reference Signal (SRS). Perfect CSI estimations are assumed here, as in this chapter, we only focus on the introduction of a low complexity partially decentralized algorithm. The inaccurate scenario is left for future work.

5.1.3 Problem Statement

For the interpretation of the partially decentralized mechanism, we select the transmission power minimization problem as the example: The network aims to minimize the *weighted* sum energy consumption of all UEs, with guaranteed achievable rate for each UE, i.e.,

$$\mathcal{P}: \min \sum_{k=1}^{K} u_k \eta_k \tag{5.3}$$

$$\text{s.t.} \quad R_k(\boldsymbol{\eta}) \geq \mathcal{R}_k, \ \forall k \in \mathcal{K} \tag{5.4}$$

$$r_{\text{FH},n}(\boldsymbol{\eta}) \leq C_{\text{FH},n}, \ \forall n \in \mathcal{N} \tag{5.5}$$

where parameter u_k is the predetermined weight factor of UE k, which, for example, can be determined by the remaining battery level of this UE. When an UE has lower battery level, its weight factor shall be set larger, so as to obtain more biased resource allocation from the network to reduce its transmit power. The power allocation vector among all UEs is denoted by $\boldsymbol{\eta} = [\eta_1, \eta_2, ..., \eta_N]^T$. $R_k(\boldsymbol{\eta})$ in (5.4) indicates the achievable rate of UE k, which is a function of the power allocation scheme, as well as the channel coefficients. $\mathcal{R}_k$ denotes the target rate. Furthermore, $r_{\text{FH},n}$ in (5.5) denotes the fronthaul capacity required for the delivery of the superposed signals from eRRH n to the BBU pool. The analytical expressions of R_k and $r_{\text{FH},n}$ depend on the decentralized mechanism we are going to propose, and will be given in next subsections.

Remark: Such a power minimization problem is particular suitable for the Offloading (Mobile Edge Computing) scenario [Mao+17], in which the tasks (e.g., VR tasks) of an UE are not executed locally. However, these tasks are offloaded via the uplink transmission to some BSs or the cloud with huge computational capabilities. Such a procedure is beneficial to UEs, as long as the energy consumed by the uplink transmission, is smaller than the energy consumed by executing the computation locally. Moreover, the rate for offloading tasks must be guaranteed, in order to ensure the overall latency at least not larger than the local execution latency.

5.2 Decentralized Approach and Algorithm

In this section, we are going to elaborate on how to solve the problem raised above in a partially decentralized manner. The conventional centralized approach will be compared with the proposed partially decentralized one.

5.2.1 The Conventional Approach

It seems that the problem $\mathcal{P}$ (5.3)-(5.5), or some related problems, shall be solved in a centralized manner, as introduced in Chapter 3, and in many existing works [Par+13a; Par+13b; Par+14; SYC14; SZL14; ZY14; LBZ15; DY16b]. This is mainly due to the fact that, for any scheduled UE, in either uplink or downlink, its achievable rate depends not only on itself, but also on all other UEs, as the uplink data streams can interfere with each other. Obviously, this is also true in our problem $\mathcal{P}$. Hence, a centralized optimization procedure is required, as long as an optimal solution is the target. Before we deep into the introduction of the new mechanism, we firstly review the signal processing procedure for the uplink of C-RAN, in which the conventional centralized approach is adopted, as shown in Fig. 5.1, where the tasks executed at each component of the network are listed.

In such a widely adopted conventional centralized approach, the instantaneous CSI knowledge from all UEs to each antenna are estimated at RRHs. The superposed analog signals at each antenna are then compressed. Hence, a compressor must be configured for each antenna. The compressed signals are then forwarded via fronthauls to the cloud. At the BBU pool, a joint decompression, detection and decoding (JDD) procedure is executed. Moreover, the whole network optimization is also performed there, with the global CSI knowledge collected from each RRH based on SRS. It has been claimed in [Par+13b; Par+14; ZY14], that such a centralized signal processing and optimization strategy is optimal, from the perspective of the information theory.

As stated before, there are several drawbacks of such a centralized mechanism:

1. Considering the delivery of the global CSI knowledge via pilots to the BBU pool, it is apparent that such a procedure introduces huge amount of overhead, as well as occupies a certain amount of the fronthaul resources, especially when there is a large number of antennas.

2. Some distortions to the CSI are inevitably introduced during such a delivery process. Hence, the BBU pool cannot obtain perfect CSI knowledge.

3. The complexity for implementing the optimization procedure scales super-exponentially with the total number of antennas.

The last item listed above prevents us from realizing a simple combination of Massive MIMO with C-RAN, i.e., replacing each eRRH with a Massive MIMO system and then applying the same centralized optimization algorithms introduced in last chapters. Even with moderate number of antennas [Par+13b], the complexity of such a JDD mechanism depicted in Fig. 5.1 has already been extremely high. Hence, such a theoretically optimal centralized strategy is **not scalable** to more antennas and more eRRHs for better performance. For example, in some existing work like [SYC14], the compression process is optimized per antenna, then for a large number of antennas, the complexity becomes unacceptable. Even if the fog computing capability is adopted, e.g., as in [Pen+16; PSS16; Tao+16], the global CSI knowledge is still requested by the BBU pool, the huge amount of overhead and the high complexity of the optimization still makes it difficult to be implemented in practice. Therefore, some new signal processing mechanisms and optimization algorithms are needed, for the practical realization of the Massive MIMO based F-RAN.

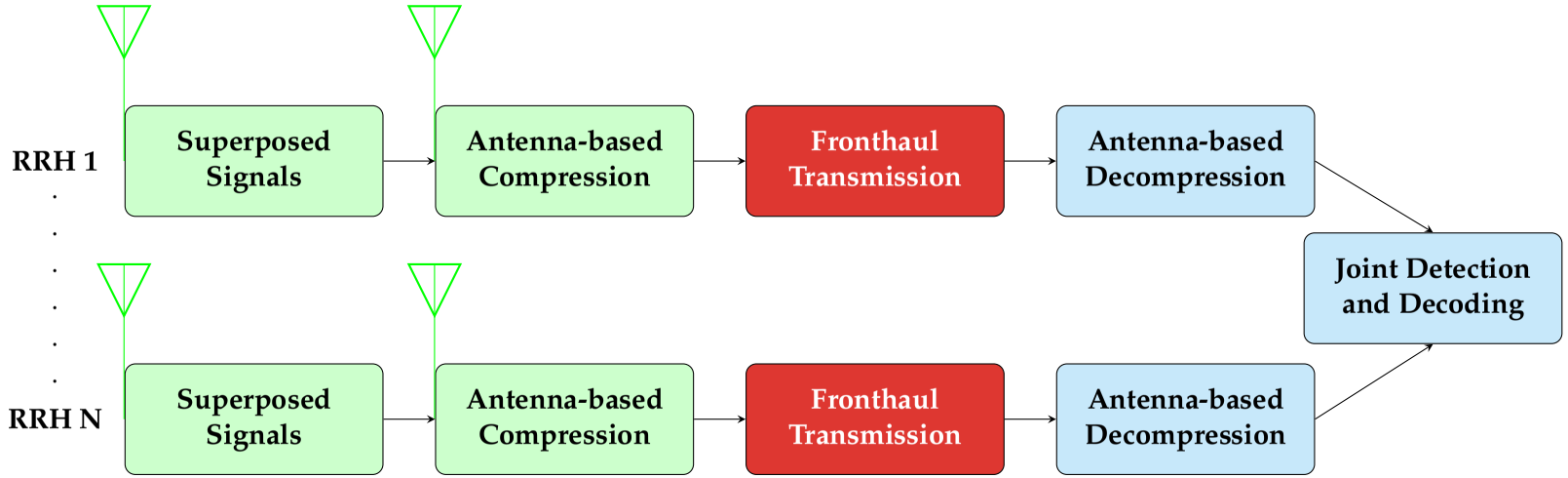


Figure 5.1: The conventional signal processing flow chart for the uplink of C-RAN. Green: at RRHs equipped with a single antenna; Red: at Fronthauls; Blue: at the BBU pool.

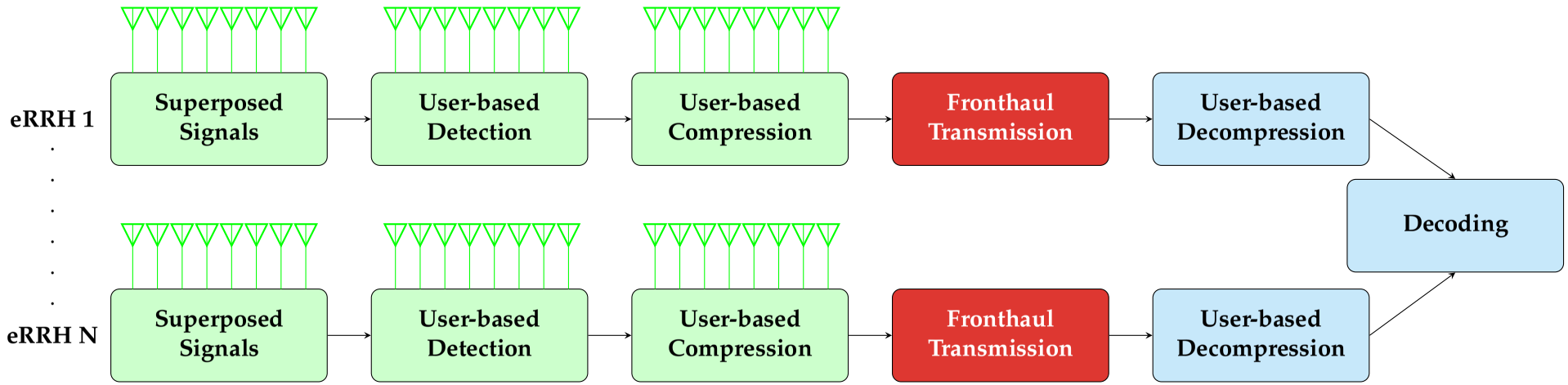


Figure 5.2: The proposed signal processing flow chart for the uplink of Massive MIMO based F-RAN. Green: at eRRHs equipped with a moderate number of antennas; Red: at Fronthauls; Blue: at the BBU pool.

5.2.2 The Proposed Approach

In this subsection, we will propose a new signal processing mechanism by adopting the fog computing capabilities of each eRRH. The overall signal processing flow chart is shown in Fig. 5.2. With this new mechanism, the amount of overhead, the complexity of the optimization, as well as the hardware costs, can be greatly reduced, especially when there are a large number of antennas. Of course everything has its price, the numerical results, which will be presented later, show that the proposed mechanism introduces some minor performance degradation. However, the much less computational complexity compared with the conventional centralized approach, can make such a minor degradation quite worthy.

From now on we describe the signal processing process for each function block in Fig. 5.2 in detail. Each subsection below corresponds to a single function block. Let's start with the very beginning on the left.

5.2.2.1 Superposed Signals Received at eRRH

After the scheduled UEs have sent their signals, at each eRRH, the received signal is a superposition of them, which are distorted independently by the channel vectors between UEs and eRRHs as well as the noise. Together with (5.1) and (5.2), the received signal vector $\mathbf{y}_n \in \mathbb{C}^{L\times 1}$ at eRRH n can be expressed the same way as in the conventional approach, i.e.,

$$\mathbf{y}_n = \mathbf{G}_n\mathbf{x} + \mathbf{z}_n = \sqrt{\rho_{\text{ul}}}\mathbf{H}_n\mathbf{D}_{\boldsymbol{\beta}_n}^{1/2}\mathbf{D}_{\boldsymbol{\eta}}^{1/2}\mathbf{s} + \mathbf{z}_n, \tag{5.6}$$

where the additive white Gaussian noise vector at eRRH n is denoted by $\mathbf{z}_n = [z_n^1, z_n^2, ..., z_n^L]^T \in \mathbb{CN}(\mathbf{0}, \sigma^2\mathbf{I}_{L\times L})$. The small scale fading matrix consisting of all small scale fading coefficients between UEs and eRRHs is constructed as $\mathbf{H}_n = [\mathbf{h}_{n,1}, \mathbf{h}_{n,2}, ..., \mathbf{h}_{n,K}] \in \mathbb{C}^{L\times K}$ with $\mathbf{h}_{n,k} = [h_{n,k}^1, h_{n,k}^2, ..., h_{n,k}^L]^T$. Similarly, the large scale fading matrix can be expressed as $\mathbf{D}_{\boldsymbol{\beta}_n} = \text{Diag}\left([\beta_{n,1}, \beta_{n,2}, ..., \beta_{n,K}]^T\right)$.

5.2.2.2 UE-based MRC Detection Process at eRRH

In contrast to the centralized mechanism, where a joint detection procedure is executed in the cloud, this new approach adopts the fog computing capabilities of the eRRHs, and pulls the detection procedure from the BBU pool back to the eRRHs. Here the Maximal Ratio Combining (MRC) detection is adopted at each eRRH for different UEs. The MRC detection has the following advantages:

1. The complexity of MRC detection is quite low. Moreover, it allows for distributed detection executed locally and independently at each eRRH. Hence, the local CSI knowledge obtained from SRS is sufficient, without the necessity of exchanging the CSI among all eRRHs and delivering them to the BBU pool. Thus, the amount of overhead can be greatly reduced.

2. For the asymptotically favorable propagation scenario with a large number of antennas, the MRC detection is close to be optimal. For details please refer to [Mar10; Mar+16],

By regarding $\mathbf{D}_{\beta_n}^{1/2}\mathbf{D}_{\boldsymbol{\eta}}^{1/2}\mathbf{s}$ in (5.6) as the *equivalent* transmit matrix from all UEs to eRRH n, the MRC detection matrix $\mathbf{D}_{\mathrm{MRC},n} \in \mathbb{C}^{L\times K}$ for eRRH n actually becomes

$$\mathbf{D}_{\mathrm{MRC},n} = \mathbf{H}_n \tag{5.7}$$

Obviously, the MRC detection matrix can be constructed locally at eRRH n by using the local CSI knowledge. Moreover, as we are going to show later, it is not necessary to deliver $\{\mathbf{H}_n\}_{n=1}^N$ to the BBU pool. Hence, as stated before, compared with the conventional approach, a huge amount of overhead can be eliminated.

After the MRC detection is executed by eRRH n, the estimated symbol $d_{n,k}$ for UE k is expressed in (5.8).

$$d_{n,k} = \left[\mathbf{D}_{\mathrm{MRC},n}^H \mathbf{y}_n\right]_k = \underbrace{\sqrt{\rho_{\mathrm{ul}}\beta_{n,k}\eta_k}||\mathbf{h}_{n,k}||^2 s_k}_{\text{desired}} + \underbrace{\mathbf{h}_{n,k}^H\left(\sum_{\substack{k'=1\\k'\neq k}}^{K}\sqrt{\rho_{\mathrm{ul}}\beta_{n,k'}\eta_{k'}}\mathbf{h}_{n,k'}s_{k'} + \mathbf{z}_n\right)}_{\text{residual interference and noise}}. \tag{5.8}$$

Obviously, in addition to the additive noise, some residual interference from the symbols of other UEs are still incorporated at each estimated symbol. According to [Mar10], when the number of antennas becomes larger and larger, such residual interference would become more and more negligible compared with the desired symbol. It should be noted that after the MRC detection procedure for each UE, there are K parallel data streams constructed at each eRRH, instead of L in the conventional centralized approach, i.e., independent of the number of antennas.

5.2.2.3 UE-based Compression Process at eRRH

Similar to the conventional approach, due to the limited fronthaul capacity, the estimated symbols from the previous process should be compressed before being fronthauled to the BBU pool. Specifically, in the proposed approach, eRRH n compresses

the estimated $\{d_{n,k}\}_{k=1}^{K}$ to $\{\widetilde{d}_{n,k}\}_{k=1}^{K}$. The compression procedure is performed per UE, instead of per antenna in the conventional approach. We also adopt the quantization to realize the compression. Hence, for each eRRH, only K quantizers are required to perform the compression, instead of scaling with the value of L. At the side of the BBU pool, only the quantized signal, i.e., $\{\widetilde{d}_{n,k}\}_{k=1}^{K}$ can be reconstructed.

For ease of analysis, we adopt the widely used (e.g. in [GK11; Par+14; ZY14; PSS16]) method to model the quantization process, i.e., it is modeled by adding artificial quantization noise to the original signal, as shown in (5.9): The quantized symbol is obtained by superposing the Gaussian distributed quantization noise $q_{n,k}$ with variance $Q_{n,k}$, which is independent of $d_{n,k}$, to the estimated symbol $d_{n,k}$. Note that we do not use the well-known Wyner-Ziv coding to model the quantization, thus the AIB method proposed in Subsection 3.2.1 will also not be adopted for the optimization, as such specific quantization schemes will complicate the analysis below. We only want to introduce a decentralized approach and show its benefit from the perspective of information theory. The quantization model adopted here is a general tool to analyse the information-theoretic performance. Any specific quantization scheme obtained via the AIB method, can be regarded as a special case of this model. According to the rate distortion theory [GK11], the compression rate $r_{n,k}$ for symbol $d_{n,k}$ can be expressed in (5.10).

$$\widetilde{d}_{n,k} = d_{n,k} + q_{n,k}, \text{ with } q_{n,k} \sim \mathbb{C}\mathcal{N}(0, Q_{n,k}) \tag{5.9}$$

$$r_{n,k} = \log_2\left(1 + \frac{\mathrm{Var}(d_{n,k})}{Q_{n,k}}\right) \tag{5.10}$$

If the fronthaul can support rate $r_{n,k}$, then the BBU pool can definitely reconstruct $\widetilde{d}_{n,k}$ via the UE-based decompression in next steps. A stronger compression for UE k at eRRH n will lead to a larger value of the variance $q_{n,k}$ and a higher distortion level, but a lower compression rate $r_{n,k}$ can be achieved.

Before directly computing the analytical expression of $\mathrm{Var}(d_{n,k})$, we reformulate (5.8) as follows

$$\begin{aligned} d_{n,k} = & \sqrt{\rho_{\mathrm{ul}}\beta_{n,k}\eta_k}\mathbb{E}\left\{||\mathbf{h}_{n,k}||^2\right\} s_k \\ & + \sqrt{\rho_{\mathrm{ul}}\beta_{n,k}\eta_k}\left(||\mathbf{h}_{n,k}||^2 - \mathbb{E}\left\{||\mathbf{h}_{n,k}||^2\right\}\right) s_k \\ & + \mathbf{h}_{n,k}^H\left(\sum_{\substack{k'=1 \\ k'\neq k}}^{K}\sqrt{\rho_{\mathrm{ul}}\beta_{n,k'}\eta_{k'}}\mathbf{h}_{n,k'}s_{k'} + \mathbf{z}_n\right). \end{aligned} \tag{5.11}$$

A similar method introduced in [Mar+16] is adopted here to compute $\mathrm{Var}(d_{n,k})$ based on (5.11): The variance of each term of $\mathrm{Var}(d_{n,k})$ will be computed separately, then they will be added together.

The first term in (5.11) denotes the desired symbol. Note that $\mathbb{E}\{s_k\} = 0$ and $\mathrm{Var}(s_k) = 1,\ \forall k$, its variance can be expressed as

$$\mathrm{Var}\left(\sqrt{\rho_{\mathrm{ul}}\beta_{n,k}\eta_k}\mathbb{E}\left\{||\mathbf{h}_{n,k}||^2\right\}s_k\right) = \rho_{\mathrm{ul}}\beta_{n,k}\eta_k\left(\mathbb{E}\left\{||\mathbf{h}_{n,k}||^2\right\}\right)^2 = L^2\rho_{\mathrm{ul}}\beta_{n,k}\eta_k, \quad (5.12)$$

by noting that the second-order moment $\mathbb{E}\left\{||\mathbf{h}_{n,k}||^2\right\} = L$.

The second term has variance

$$\begin{aligned}&\mathrm{Var}\left(\sqrt{\rho_{\mathrm{ul}}\beta_{n,k}\eta_k}\left(||\mathbf{h}_{n,k}||^2 - \mathbb{E}\left\{||\mathbf{h}_{n,k}||^2\right\}\right)s_k\right)\\ &= \rho_{\mathrm{ul}}\beta_{n,k}\eta_k\left(\mathbb{E}\left\{||\mathbf{h}_{n,k}||^4\right\} - \left(\mathbb{E}\left\{||\mathbf{h}_{n,k}||^2\right\}\right)^2\right)\\ &= L\rho_{\mathrm{ul}}\beta_{n,k}\eta_k, \end{aligned} \quad (5.13)$$

by noting that the fourth-order moment $\mathbb{E}\left\{||\mathbf{h}_{n,k}||^4\right\} = L(L+1)$.

The third term denotes the interference from the non-orthogonality of the channel and the additive noise, it has variance

$$\begin{aligned}&\mathrm{Var}\left(\mathbf{h}_{n,k}^H\left(\sum_{\substack{k'=1\\k'\neq k}}^{K}\sqrt{\rho_{\mathrm{ul}}\beta_{n,k'}\eta_{k'}}\mathbf{h}_{n,k'}s_{k'} + \mathbf{z}_n\right)\right)\\ &= \rho_{\mathrm{ul}}\sum_{\substack{k'=1\\k'\neq k}}^{K}\beta_{n,k'}\eta_{k'}\mathbb{E}\left\{||\mathbf{h}_{n,k}||^2\right\} + \mathbb{E}\left\{||\mathbf{z}_n||^2\right\}\mathbb{E}\left\{||\mathbf{h}_{n,k}||^2\right\}\\ &= L\rho_{\mathrm{ul}}\sum_{\substack{k'=1\\k'\neq k}}^{K}\beta_{n,k'}\eta_{k'} + \sigma^2 L. \end{aligned} \quad (5.14)$$

Therefore, the analytical expression of $\mathrm{Var}(d_{n,k})$ can be expressed as

$$\begin{aligned}\mathrm{Var}(d_{n,k}) &= L^2\rho_{\mathrm{ul}}\beta_{n,k}\eta_k + L\rho_{\mathrm{ul}}\beta_{n,k}\eta_k + L\rho_{\mathrm{ul}}\sum_{\substack{k'=1\\k'\neq k}}^{K}\beta_{n,k'}\eta_{k'} + \sigma^2 L\\ &= L^2\rho_{\mathrm{ul}}\beta_{n,k}\eta_k + L\rho_{\mathrm{ul}}\sum_{\substack{k'=1\\k'\neq k}}^{K}\beta_{n,k'}\eta_{k'} + \sigma^2 L. \end{aligned} \quad (5.15)$$

Similar to the conventional centralized approach, the compression process in this scheme is also determined by the value of $Q_{n,k}\ \forall n,k$. Thus it is also subject to be optimized. Note that with this new approach, the number of variables and constraints are linearly dependent on K, instead of L. In scenarios when $K \ll L$, it has much lower complexity than the conventional antenna-based approach.

5.2.2.4 Fronthauling from eRRHs to the BBU pool

As stated before, there are K data streams compressed at each eRRH. For eRRH n, the compression rate $r_{\mathrm{FH},n}$ of the compressed signals can be expressed as:

$$r_{\mathrm{FH},n} = \sum_{k=1}^{K} r_{n,k} = \sum_{k=1}^{K} \log_2 \left(1 + \frac{L^2 \rho_{\mathrm{ul}} \beta_{n,k} \eta_k + L \rho_{\mathrm{ul}} \sum_{\substack{k'=1 \\ k' \neq k}}^{K} \beta_{n,k'} \eta_{k'} + \sigma^2 L}{Q_{n,k}} \right). \quad (5.16)$$

5.2.2.5 UE-based Reconstruction at the BBU pool

According to the rate distortion theory [GK11], which is also introduced in Section 2.1, the quantized signal from eRRH n, i.e., $\{\widetilde{d}_{n,k}\}_{k=1}^{K}$, can be reconstructed at the BBU pool, as long as $r_{\mathrm{FH},n} \leq C_{\mathrm{FH},n}$ is fulfilled.

5.2.2.6 Decoding Process at the BBU pool

We see that the signal from the same UE is independently received and compressed at all eRRHs. Therefore, after the decompression and reconstruction procedure executed at the BBU pool, the signals from the same UE can be combined. Specifically, by adding the reconstructed signals of UE k from all eRRHs together, i.e., $\{\widetilde{d}_{n,k}\}_{n=1}^{N}\ \forall n$, we obtain

$$\begin{aligned} \widetilde{d}_k = \sum_{n=1}^{N} \widetilde{d}_{n,k} = & \underbrace{\sum_{n=1}^{N} \sqrt{\rho_{\mathrm{ul}} \beta_{n,k} \eta_k} ||\mathbf{h}_{n,k}||^2 s_k}_{\text{desired}} \\ & + \underbrace{\sum_{n=1}^{N} \mathbf{h}_{n,k}^H \left(\sum_{\substack{k'=1 \\ k' \neq k}}^{K} \sqrt{\rho_{\mathrm{ul}} \beta_{n,k'} \eta_{k'}} \mathbf{h}_{n,k'} s_{k'} + \mathbf{z}_n \right) + \sum_{n=1}^{N} q_{n,k}}_{\text{residual interference and noise}}. \end{aligned} \quad (5.17)$$

From (5.17), we see that the signal $\widetilde{d}_k$ incorporates both the desired part gathered from all eRRHs, as well as the residual interference, the additive noise, and the quantization noise introduced by the compression procedure. By treating all interference as noise, the BBU pool can decode s_k from $\widetilde{d}_k$. Again, by adopting the technique when (5.15) is computed, the achievable SINR for UE k, as well as the

corresponding achievable rate can be computed as follows:

$$\begin{aligned}\mathrm{SINR}_k &= \frac{\left(\sum_{n=1}^{N} L\sqrt{\rho_{\mathrm{ul}}\beta_{n,k}\eta_k}\right)^2}{\sum_{n=1}^{N}\left(L\rho_{\mathrm{ul}}\sum_{\substack{k'=1\\k'\neq k}}^{K}\beta_{n,k'}\eta_{k'}+\sigma^2 L+Q_{n,k}\right)} \\ &= \frac{\left(\sum_{n=1}^{N} L\sqrt{\rho_{\mathrm{ul}}\beta_{n,k}}\right)^2\eta_k}{\sum_{n=1}^{N}\left(L\rho_{\mathrm{ul}}\sum_{\substack{k'=1\\k'\neq k}}^{K}\beta_{n,k'}\eta_{k'}+\sigma^2 L+Q_{n,k}\right)}, \end{aligned} \tag{5.18}$$

$$R_k = \log_2\left(1+\mathrm{SINR}_k\right). \tag{5.19}$$

5.2.3 Final Problem Formulation and Solution

With the analytical expressions derived above, original problem $\mathcal{P}$ (5.3)-(5.5) can be reformulated as follows

$$\mathcal{P}: \min_{\{\eta_k\}_{k=1}^K,\{Q_{n,k}\}_{n=1,k=1}^{N,K}} \sum_{k=1}^{K} u_k\eta_k, \tag{5.20}$$

$$\text{s.t.} \quad \log_2\left(1+\mathrm{SINR}_k\right) \geq \mathcal{R}_k, \quad \forall k\in\mathcal{K}, \tag{5.21}$$

$$r_{\mathrm{FH},n} \leq C_{\mathrm{FH},n}, \quad \forall n\in\mathcal{N} \tag{5.22}$$

$$0\leq\eta_k\leq 1,\ Q_{n,k}\geq 0, \quad \forall k\in\mathcal{K},\ \forall n\in\mathcal{N}. \tag{5.23}$$

The objective (5.20) is linear with respect to $\{\eta_k\}_{k=1}^K$, and thus convex. And constraints (5.23) are also convex.

For (5.21), it can be equivalently expressed as (5.24). Obviously, it is also linear with respect to $\{\eta_k\}_{k=1}^K$ and $\{Q_{n,k}\}_{n=1,k=1}^{N,K}$ and thus convex.

$$(2^{\mathcal{R}_k}-1)\left(L\rho_{\mathrm{ul}}\sum_{\substack{k'=1\\k'\neq k}}^{K}\beta_{n,k'}\eta_{k'}+\sigma^2 L+Q_{n,k}\right)-\left(\sum_{n=1}^{N} L\sqrt{\rho_{\mathrm{ul}}\beta_{n,k}}\right)^2\eta_k\leq 0, \quad \forall k\in\mathcal{K}. \tag{5.24}$$

However, the LHS of (5.22) is still not convex. Fortunately, we can adopt the same iterative approximation method for convexification, which has already been derived and used in the previous chapter: By introducing auxiliary variables $\{\ell_{n,k}\}$, the LHS of (5.22) can be expressed as (5.25), and it is upper-bounded by (5.26). For known values of $\{\ell_{n,k}\}$, (5.26) is a linear with respect to $\{\eta_k\}$ and $\{Q_{n,k}\}$, and thus convex. Hence, the LHS of (5.22) can be approximated iteratively by (5.26). At the start of each iteration, the value of $\{\ell_{n,k}\}$ will be updated according to (5.27), based on the results from the previous iteration. For more details please refer to the derivations (4.51)-(4.60) in Subsection 4.2.1.2.

$$\sum_{k=1}^{K}\log_2\left(Q_{n,k}+L^2\rho_{\mathrm{ul}}\beta_{n,k}\eta_k+L\rho_{\mathrm{ul}}\sum_{\substack{k'=1\\k'\neq k}}^{K}\beta_{n,k'}\eta_{k'}+\sigma^2L\right)-\sum_{k=1}^{K}\log_2\left(Q_{n,k}\right) \tag{5.25}$$

$$\leq\sum_{k=1}^{K}\left(\log_2\ell_{n,k}+\frac{Q_{n,k}+L^2\rho_{\mathrm{ul}}\beta_{n,k}\eta_k+L\rho_{\mathrm{ul}}\sum_{\substack{k'=1\\k'\neq k}}^{K}\beta_{n,k'}\eta_{k'}+\sigma^2L}{\ell_{n,k}\ln 2}\right)$$

$$-\frac{K}{\ln 2}-\sum_{k=1}^{K}\log_2\left(Q_{n,k}\right), \tag{5.26}$$

the equality holds when $\ell_{n,k}=Q_{n,k}+L^2\rho_{\mathrm{ul}}\beta_{n,k}\eta_k+L\rho_{\mathrm{ul}}\sum_{\substack{k'=1\\k'\neq k}}^{K}\beta_{n,k'}\eta_{k'}+\sigma^2L.$

(5.27)

Therefore, with the techniques introduced above, problem (5.20)-(5.23) can be further approximated as the following convex problem. For the $(t+1)$-th iteration, the problem can be written as

$$\mathcal{P}^{(t+1)}:\min_{\{\eta_k^{(t+1)}\}_{k=1}^{K},\{Q_{n,k}^{(t+1)}\}_{n=1,k=1}^{N,K}}\quad\sum_{k=1}^{K}u_k\eta_k^{(t+1)}, \tag{5.28}$$

$$\text{s.t.}\quad(2^{\mathcal{R}_k}-1)\left(L\rho_{\mathrm{ul}}\sum_{\substack{k'=1\\k'\neq k}}^{K}\beta_{n,k'}\eta_{k'}^{(t+1)}+\sigma^2L+Q_{n,k}^{(t+1)}\right)$$

$$-\left(\sum_{n=1}^{N}L\sqrt{\rho_{\mathrm{ul}}\beta_{n,k}}\right)^2\eta_k^{(t+1)}\leq 0,\quad\forall k\in\mathcal{K}, \tag{5.29}$$

$$\sum_{k=1}^{K}\left(\log_2\ell_{n,k}^{(t+1)}+\frac{Q_{n,k}^{(t+1)}+L^2\rho_{\mathrm{ul}}\beta_{n,k}\eta_k^{(t+1)}+L\rho_{\mathrm{ul}}\sum_{\substack{k'=1\\k'\neq k}}^{K}\beta_{n,k'}\eta_{k'}^{(t+1)}+\sigma^2L}{\ell_{n,k}^{(t+1)}\ln 2}\right)$$

$$-\frac{K}{\ln 2}-\sum_{k=1}^{K}\log_2\left(Q_{n,k}^{(t+1)}\right)-C_{\mathrm{FH},n}\leq 0,\quad\forall n\in\mathcal{N}, \tag{5.30}$$

$$0\leq\eta_k^{(t+1)}\leq 1,\ Q_{n,k}^{(t+1)}\geq 0,\quad\forall k\in\mathcal{K},\ \forall n\in\mathcal{N}. \tag{5.31}$$

The auxiliary parameter $\ell_{n,k}^{(t+1)}$ is calculate from the results from the previous iteration, i.e.,

$$\ell_{n,k}^{(t+1)}=Q_{n,k}^{(t)}+L^2\rho_{\mathrm{ul}}\beta_{n,k}\eta_k^{(t)}+L\rho_{\mathrm{ul}}\sum_{\substack{k'=1\\k'\neq k}}^{K}\beta_{n,k'}\eta_{k'}^{(t)}+\sigma^2L,\quad\forall k\in\mathcal{K},\ \forall n\in\mathcal{N}. \tag{5.32}$$

In order to obtain the initial value of $\ell_{n,k}$, constraint (5.22) is temporarily dropped

to form the initial problem:

$$\mathcal{P}^{(0)}: \min_{\{\eta_k^{(0)}\}_{k=1}^K, \{Q_{n,k}^{(0)}\}_{n=1,k=1}^{N,K}} \quad \sum_{k=1}^{K} u_k \eta_k^{(0)}, \tag{5.33}$$

$$\text{s.t.} \quad (2^{\mathcal{R}_k}-1)\left(L\rho_{\text{ul}} \sum_{\substack{k'=1 \\ k'\neq k}}^{K} \beta_{n,k'} \eta_{k'}^{(0)} + \sigma^2 L + Q_{n,k}^{(0)}\right) - \left(\sum_{n=1}^{N} L\sqrt{\rho_{\text{ul}}\beta_{n,k}}\right)^2 \eta_k^{(0)} \leq 0, \quad \forall k \in \mathcal{K}, \tag{5.34}$$

$$0 \leq \eta_k^{(0)} \leq 1, \; Q_{n,k}^{(0)} \geq 0, \quad \forall k \in \mathcal{K}, \; \forall n \in \mathcal{N}. \tag{5.35}$$

Obviously, with the proposed signal processing mechanism, the original problem is reformulated and approximated as a convex optimization problem, whose most constraints are linear. By solving this problem, the power control and the compression procedure are optimized jointly. We summarize the solving steps of the problem in Alg. 8:

Algorithm 8: The iterative optimization steps for the uplink of Massive MIMO based F-RAN

1 **Initialization**: Construct and solve the Linear Programming (LP) initial problem $\mathcal{P}^{(0)}$ according to (5.33)-(5.35), base on the solutions the initial values of $\{\ell_{n,k}^{(1)}\}_{n=1,k=1}^{N,K}$ can be obtained according to (5.32). Construct the problem $\mathcal{P}^{(1)}$ according to (5.28)-(5.31), and set $t \leftarrow 1$.
2 **repeat**
3 Solve the problem $\mathcal{P}^{(t)}$.
4 Compute the values of $\{\ell_{n,k}^{(t+1)}\}_{n=1,k=1}^{N,K}$ based on (5.32).
5 Formulate the problem $\mathcal{P}^{(t+1)}$ according to (5.28)-(5.31), and set $t \leftarrow t+1$.
6 **until** *Convergence or reaching max iteration number*;

5.2.4 Comparison with the Conventional Centralized Approach

In order to demonstrate the benefits of the proposed approach, in this subsection, we give a brief comparison between the proposed approach (**P**), and the conventional centralized approach (**C**):

Alg.	Overhead	Compression	Complexity
P	$K\mathcal{Q}/T_{\boldsymbol{\beta}}$	K parallel	$\mathcal{O}(K^4)$
C	$L\mathcal{Q}/T_{\boldsymbol{h}}$	L parallel	$\mathcal{O}(L^4)$

1. ***Overhead*:** As analyzed in the previous subsection, with the proposed approach, the global CSI knowledge is not required at the BBU pool anymore, only the large scale fading coefficients $\boldsymbol{\beta} = \{\beta_{n,k}\}_{n,k=1}^{N,K}$ need to be delivered. The small scale fading coefficients $\boldsymbol{h} = \{h_{n,k}^{l}\}_{n,k,l=1}^{N,K,L}$ are collected and used locally at the distributed eRRHs, for executing the MRC detection. Let $\mathcal{Q}$ be the number of bits required to describe the CSI, and $T_{\boldsymbol{\beta}}$ be the duration that the large scale fading coefficients stay unchanged, then we can approximate the bits of overhead delivered from each eRRH to the BBU pool as $K\mathcal{Q}/T_{\boldsymbol{\beta}}$ per second. However, in the conventional centralize approach, the BBU pool have to execute all signal processing functionalities, thus the instantaneous CSI between each UE and each antenna, must be available at the BBU pool in the cloud. Hence, the corresponding overhead can be approximated as $L\mathcal{Q}/T_{\boldsymbol{h}}$ per second. In general, the small scale fading coefficients vary much faster than the large scale fading coefficients, i.e., we usually have $T_{\boldsymbol{\beta}} \gg T_{\boldsymbol{h}}$.

2. ***Compression*:** As stated before, due to the UE-based detection in our proposed approach, there are K parallel data streams needed to be compressed. Hence, only K quantizers are required at each eRRH, instead of L in the antenna-based compression with the conventional approach.

3. ***Complexity*:** The complexity for solving the problem $\mathcal{P}$ is $\mathcal{O}(K^4)$, instead of $\mathcal{O}(L^4)$ with the centralized antenna-based approach.

From such a brief comparison we see that, by utilizing the fog computing capabilities at eRRHs, with which many signal processing functionalities can be executed at distributed eRRHs in a decentralized way, the amount of overhead, the number of quantizers as well as the complexity of the overall algorithm will scale with the number of the scheduled UEs, i.e., K, instead of scaling with the number of antennas among eRRHs, i.e., L, in the conventional centralized approach. Hence, increasing the number of antennas L will not increase the complexity, the amount of the overhead, as well as the latency and the hardware cost related to the compression. Such a property makes a scalable architecture of the Massive MIMO based F-RAN possible, i.e., the network providers can simply equip the eRRH with more antennas for better performance, but without the need to worry about higher complexity and cost.

We must emphasize that in the proposed approach, all analytical expressions and derivations above are valid for arbitrary values of L. Similar to Massive MIMO, the performance approaches the theoretical limits only when L is sufficient large, by virtue of more effective channel hardening effect[Mar+16].

Table 5.1: The simulation parameters for Massive MIMO based F-RAN.

Scenario	Dense urban
Cell radius r_{cell}	0.5 km
Number of eRRHs N	7
Number of antennas per eRRH L	32, 64, 128, 256
Fronthaul capacity C_{FH}	1.5 Gbps
Total number of UE: K_{total}	64
Number of scheduled UEs per UL slot K	32
Maximal transmission power per UE	20 dBm
Carrier frequency f_c	1.9 GHz
Network bandwidth B	20 MHz
Uplink rate target per UE $\mathcal{R}_{\text{pu}}$	20, 40, 50 Mbps
eRRH antenna height	30 m
UE antenna height	1.5 m
Wireless path loss model	COST Hata[AA16]
Shadow fading standard deviation	8 dB
Noise temperature	300 K
eRRH receiver noise figure	9 dB
UE antenna gain	6 dBi
eRRH antenna gain	0 dBi

5.3 Numerical Results

In this section, the proposed partially decentralized mechanism will be tested via simulations. A single-cell dense urban scenario [Mar+16] is considered, and all simulation parameters are listed in Table 5.3. In the simulation scenario, seven eRRHs are positioned in the cell, each of them are equipped with L antennas to realize a networked Massive MIMO F-RAN system. An eRRH is mounted at the center of the cell, and the other six eRRHs are uniformly positioned on the circle with radius $r_{\text{cell}}/\sqrt{2}$. There are 64 UEs randomly distributed within the cell. In each time slot, half of them, i.e., 32 UEs are scheduled to upload their tasks to the cloud server for remote computing. The scheduled UEs would like to experience guaranteed QoS, but with as less energy consumption as possible. For simplicity, we set the weight factors in (5.3) as $u_k = 1\ \forall k$. At each eRRH and the BBU pool, the proposed partially decentralized approach is implemented for signal processing and network implementation. The results will be compared with the conventional approach. In our simulation, 200 independent random realizations are set up, each of them is with random and independent UE positions and shadow fading profiles.

At first we would like to know, how much performance degradation is introduced by the proposed partially decentralized approach, with only partial CSI knowledge, compared with the centralized joint optimization with full knowledge of global CSI [Par+13b; Par+14], which is theoretically optimal. Hence, we set up the same UE

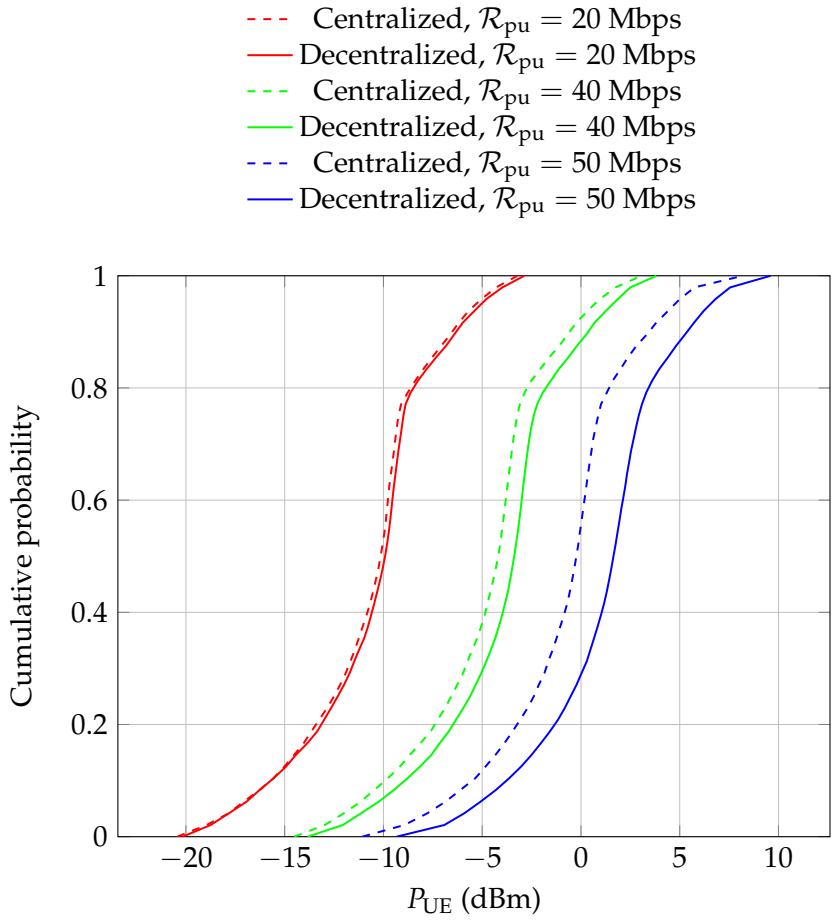


Figure 5.3: The comparison of the CDFs for the power consumption between the centralized approach and the proposed decentralized approach ($L = 128$).

target rate, as well as the same available power for 200 independent realizations. For each realization, the results of both schemes are documented after the corresponding algorithms are executed. Then the Cumulative Distribution Functions (CDF) [2] of the optimized power consumption for UEs between the two schemes are compared. The results are shown in Fig 5.3.

From the results we observe only minor degradation i.e., more power is consumed by the scheduled UEs in average. Such a degradation mainly comes from two aspects: 1. The MRC detection is performed at the eRRHs in a distributed manner, with only local CSI knowledge. Compared with the optimal joint detection and decoding [Par+14] process executed at the BBU pool, the performance loss is thus inevitable; 2. The UE-based compression process is less efficient, compared with the antenna-based compression together with the joint decompression process. Note that after executing the MRC detection at each eRRH, some residual interference (see (5.8)) is still contained at each estimated symbol. However, such interference is also compressed and thus consumes some the fronthaul resources. When higher

[2] CDF of a real-valued random variable X, or just distribution function of X, evaluated at x, is the probability that X will take a value less than or equal to x [DFO20].

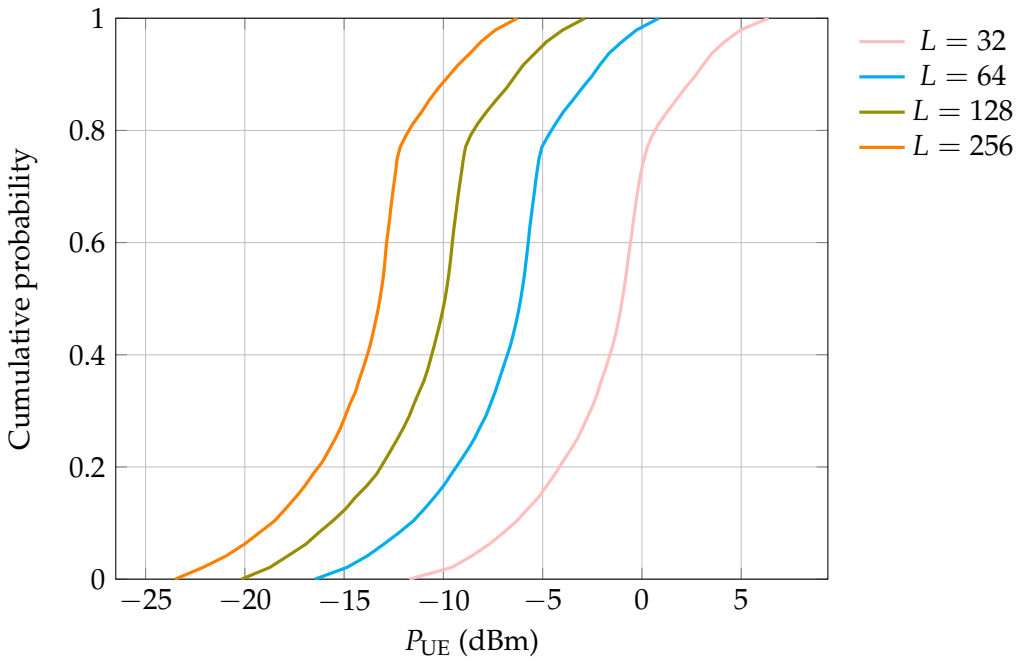


Figure 5.4: The performance comparison between different number of antennas for $\mathcal{R}_{\mathrm{pu}} = 20$ Mbps.

$\mathcal{R}_{\mathrm{pu}}$ is configured, the performance becomes more and more limited by the fronthaul resources, the advantage of the conventional approach is then more prominent, due to its more efficient utilization of the fronthaul resources.

Next, the influence of the number of antennas equipped at each eRRH is investigated. Fig. 5.4 - Fig. 5.6 illustrate the comparison of the results between different number of antennas for different per user target rates. For lower target rate $\mathcal{R}_{\mathrm{pu}}$, the power consumption can be reduced to more or less the same extent, when the number of antennas is doubled, as shown in Fig. 5.4. For higher $\mathcal{R}_{\mathrm{pu}}$, more antennas are needed so as to guarantee the target QoS. For instance, when $\mathcal{R}_{\mathrm{pu}} = 40$ Mbps and $L = 32$ antennas are at each eRRH, even the maximal allowable power of UEs cannot support their required QoS anymore. Therefore, no corresponding curve for $L = 32$ in Fig. 5.5. Moreover, the *saturation effect* can be observed for higher $\mathcal{R}_{\mathrm{pu}}$ when doubling the number of antennas, i.e., doubling the number of antennas cannot achieve the same extent of the performance improvement, which is the case when the target rate is lower. This is mainly due to the limitations resulting from the fronthaul capacity, when a higher target rate is required. It is worth to emphasize here again that with the proposed mechanism, the amount of overhead, the computational complexity, the hardware cost relating to the compression, etc., still remain unchanged when the number of antennas is doubled, as they scale only with the number of the scheduled UEs K. However, in the conventional centralized approach, it is rather difficult and expensive to double the number of antennas for achieving better performance, as they scale with the number of antennas L, as stated in Subsection 5.2.4.

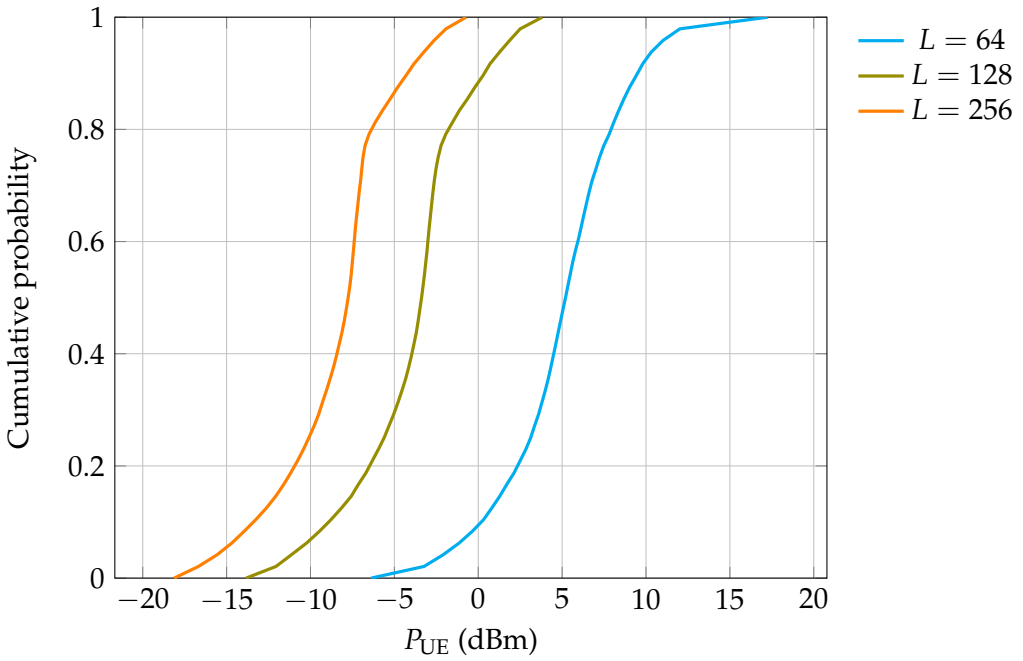


Figure 5.5: The performance comparison between different number of antennas for $\mathcal{R}_{\text{pu}} = 40$ Mbps.

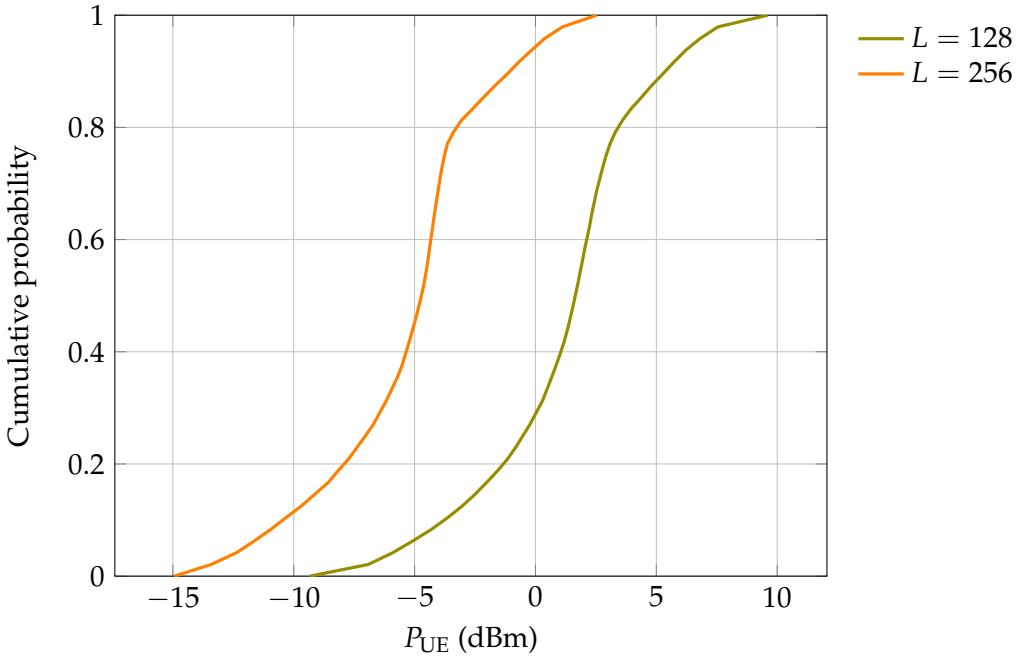


Figure 5.6: The performance comparison between different number of antennas for $\mathcal{R}_{\text{pu}} = 50$ Mbps.

5.4 Discussions, Summaries, and Outlooks

In this chapter, we give a trial on developing a scheme with which the heavy computational burden on the BBU pool can be relieved. As seen from Chapter 3 and Chapter 4, all proposed algorithms, as well as most existing schemes, require a centralized optimization step. Apart from the computational burden imposed on the BBU pool, the global CSI knowledge is also required to be available. Thus the large amount of the overhead, as well as the introduced latency, might impair the actual performance of the network in practice. The key proposed in this chapter to overcome such difficulties, is the combination of the concept of Massive MIMO, and the fog computing. As Massive MIMO can realize the channel hardening effect with the law of large numbers, with which the channel between UE and eRRHs can be regarded as a scalar deterministic channel. Hence, when Massive MIMO is combined with F-RAN, the instantaneous global CSI is not required by the BBU pool anymore, it needs only the information related to such hardened scalar channels. As a result, the amount of the overhead that convey the CSI can be greatly reduced. Moreover, as the F-RAN is actually a networked MIMO structure, such a combination can also relieve the demands on the number of antennas for a single Massive MIMO, especially when more eRRHs exist in F-RAN.

The proposed scheme consists of a decentralized signal processing mechanism executed at eRRHs, and a centralized optimization algorithm. Hence, we name it as a partially decentralized approach. The signal detection and estimation, instantaneous CSI acquisition, as well as the processing of the estimated signal are all executed locally at each eRRHs in a decentralized manner, with the help of their fog computing capabilities. By exploiting the benefit of the channel hardening effect, the BBU pool can implement a centralized optimization but with much lower computational complexity. Therefore, the combination of these two hot 5G techniques, has the potential to overcome their own limitations, and boost the performance to each other.

Although such a combination seems to be rather promising, further issues need to be analysed but cannot be addressed here. As already stated at the beginning of this chapter, the purpose of this chapter is just to pave a new way for future work. Our proposed scheme considers only the uplink of the Massive MIMO based F-RAN, targeting at the minimization of the (weighted) sum power consumption, with perfect CSI estimation at each eRRH. Hence, some interesting research directions for future are straightforward:

1. The development of a similar partially decentralized scheme for high Spectral Efficiency (SE) oriented design in the uplink, aiming to maximize the (weighted) sum rate with the power budget of each UE;

2. For high SE oriented design, wMMF between uplink UEs is also worth to be investigated, but is achieved in a decentralized manner;

3. The development of similar partially decentralized schemes, for both high SE and EE oriented design for the downlink of the Massive MIMO based F-RAN. but with performance comparable to the centralized approach proposed in Chapter 4;

4. How inaccurate CSI influences the algorithm? Is robust design possible?

5. With the proposed algorithm, a centralized optimization still needs to be executed by the BBU pool, but with less computational complexity and less requirements on the acquisition of the CSI knowledge. Is it possible to achieve a fully distributed mechanism? If possible, how about the performance degradation?

Chapter 6

Conclusions

This dissertation is a summary of our research results, in terms of how to efficiently utilize the fog computing capability in the Fog Radio Access Network (F-RAN), which is an evolution of the Cloud Radio Access Network (C-RAN). For C-RAN, nearly all computation and signal processing procedures are executed at the BBU pool in the cloud, the Remote Radio Head (RRH) just acts as Access Point (AP) of the network without almost any computation capability. While with the fog computing, the RRH evolves into the so-called enhanced RRH (eRRH), such that it can also perform some light computation tasks, as well as has the storage capability. The resultant network structure, as well as the signal processing techniques and its optimization, are the main topics studied in this work.

In Chapter 1, we have illustrated the basic idea of the network, covering the tasks of each components. Then in Chapter 2, we introduced all mathematical tools, and information theories that are required in the coming chapters that investigate F-RAN.

The trunk of the story started from Chapter 3. In this chapter, the uplink is investigated. In the uplink of both C-RAN and F-RAN, one core problem is the signal processing procedure at RRH/eRRH, i.e., how to efficiently compress the superposed signals from all scheduled UEs, such that the delivery of the compressed signal to the BBU pool, can be supported by the fronthaul and the performance, e.g., the achievable rate, is maximized. Our main contribution is the introduction of a practical quantization scheme, that can work for arbitrary codebooks, instead of only the Gaussian codebook from the perspective of information theory. The proposed quantization scheme is derived, via the execution of the proposed Alternating Information Bottleneck (AIB) method, and the alternating Bi-Section method by the BBU

pool. Furthermore, with the Outer Linearization method, the optimal fronthaul resource allocation can be obtained, when it has to be shared among RRHs/eRRHs of the network.

We put more emphasis on investigating the downlink, as in general, there are much more downlink slots than the uplink slots. The downlink performance contributes to the overall performance of the network in greater measure. As the 5G network consists of densely deployed micro Base Stations (BS), the energy consumption of a BS becomes more significant from the perspective of the network provider. Hence, we have proposed several algorithms, with which the clustering pattern of RRHs/eRRHs, the load balancing between fronthauls, and the transmission power of each RRH/eRRH can be jointly optimized, in order to minimize the transmission power under the condition that the QoS of each UE can be satisfied. The algorithms can cover different network configurations such as whether the hard or the soft transfer mode is adopted on fronthauls, or whether the fronthaul resources are dedicated to each eRRH or not. When the fronthaul resources are not dedicated, the algorithm can also provide the optimal resource allocation scheme. Based on the proposed algorithms, we also demonstrated that equipping RRHs with low-cost cache modules is a rather cheap and easy way to realize a specific type of fog computing, which can greatly improve the network performance. With caching, the cached contents can be transmitted by all eRRHs simultaneously, thus much more concentrated beams can be formed to serve UEs, which can potentially lower the power consumption and reduce the interference to others. Moreover, caching can greatly reduce the burden on fronthauls as less contents are to be conveyed by them. Hence, the available fronthaul resources can be saved for delivering other uncached contents to more eRRHs, which can further reduce the network power consumption, or increase the achievable rates.

Another significant contribution for the downlink transmission is the introduction of an algorithm, with which several eRRHs have the possibility to be switched off to save more power, and the remaining ones can still fulfill the network requirements. The eRRH deactivation has been shown to be completed within several iterations of the proposed algorithm.

When concerning the high SE oriented design, two different metrics have been studied: Multi-cast throughput maximization and weighted Max-Min Fairness. The former one is able to completely exploit the network resources but might be unfair to UEs staying at cell edges or with bad channel qualities. While the latter one can achieve the fairness, but with the price of lower throughput. Our proposed algorithms cover both scenarios. Once more, the cache module has shown its potential to increase the SE of the network in a low-cost way.

The robust design is also a contribution of this work. When only inaccurate CSI is available, with the proposed algorithm, the network achieve a robust performance: As long as the CSI inaccuracy can be bounded to some extent, even in the worst case, the network QoS can still be guaranteed. However, we have shown that the price for such robustness is higher power consumption. But the good news is that, as we have also shown, introducing the cache module can help to improve the robustness and ease the power requirements.

At last, we have a trial on developing a scheme, such that the heavy computational burden at the BBU pool, which is imposed by all proposed algorithms, can be relieved. We take an initial step by combining the technique of Massive MIMO and F-RAN. In the proposed approach, the fog computing is fully exploited to achieve some decentralized operations at eRRHs, with which less computational requirements on the BBU pool, less amount of overhead, less hardware costs, as well as shorter latency can be achieved.

Although many aspects have been covered in terms of the optimal design for F-RAN, there are always more blanks to be filled. In the last subsection of each chapter, we always listed some interesting topics that we have not solved or intensively investigated, and are worth to be studied further. With this dissertation, we would like to show some charming aspects of F-RAN, as well as to provide some design guidelines. Above all, we hope that it can help to shed some lights on possible research directions in the future.

List of Figures

List of Tables

Bibliography

[3GP18] 3GPP. *5G; NR; Physical channels and modulation (3GPP TS 38.211 version 15.2.0 Release 15)*. Project. 3GPP, 2018 (cit. on pp. 41, 79, 80, 91, 165).

[5G-15] 5G-PPP. *METIS-II*. Project. 5G-PPP, 2015 (cit. on p. 2).

[AA16] R. V. Akhpashev and A. V. Andreev. "COST 231 Hata Adaptation Model for Urban Conditions in LTE Networks". In: *International Conference of Young Specialists on Micro/Nanotechnologies and Electron Devices (EDM)* (Aug 2016) (cit. on p. 183).

[Ae11] G. Auer and *et al.* "How Much Energy is Needed to Run a Wireless Network?" In: *IEEE Wireless Communications* 18.5 (Oct 2011) (cit. on p. 82).

[Ara+17] G. Araniti, M. Condoluci, P. Scopelliti, A. Molinaro, and A. Iera. "Multicasting over Emerging 5G Networks: Challenges and Perspectives". In: *IEEE Network* 31.2 (Feb 2017) (cit. on p. 11).

[AS16] A. Alameer and A. Sezgin. "Joint Beamforming and Network Topology Optimization of Green Cloud Radio Access Networks". In: *9th International Symposium on Turbo Codes and Iterative Information Processing (ISTC)* (Sep 2016) (cit. on p. 10).

[AT12] AT and T. *Network Function Virtualization: An Introduction, Benefits, Enablers, Challenges and Call for Action*. White Paper. Oct. 2012 (cit. on p. 3).

[Ber15] D. P. Bertsekas. *Convex Optimization Algorithms*. Athena Scientific, 2015 (cit. on p. 34).

[Bon+14] F. Bonomi, R. Milito, P. Natarajan, and J. Zhu. "Fog Computing: A Platform for Internet of Things and Analytics". In: *Big Data and Internet of Things: A Roadmap for Smart Environments* (2014) (cit. on p. 11).

[BSS16] M. S. Bazaara, H. D. Sherali, and C. M. Shetty. "Nonlinear Programming". In: *Hoboken: John Wiley Sons, Inc.* (2016) (cit. on pp. 58, 62).

[BV04] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004 (cit. on pp. 31, 32, 34, 35, 140, 143).

[BW06] H. Boche and M. Wiczanowski. "Stability Optimal Transmission Policy for the Multiple Antenna Multiple Access Channel in the Geometric View". In: *EURASIP Signal Processing Journal* (Aug 2006) (cit. on p. 59).

[CCO14] D. Christopoulos, S. Chatzinotas, and B. Ottersten. "Sum Rate Maximizing Multigroup Multicast Beamforming under Per-antenna Power Constraints". In: *IEEE Global Communications Conference (GLOBECOM)* (Dec 2014) (cit. on pp. 138, 143).

[CG79] T. Cover and A. E. Gamal. "Capacity Theorems for the Relay Channel". In: *IEEE Trans. Inf. Theory* 25.5 (Sep 1979) (cit. on p. 42).

[Che+16] Z. Chen, J. Lee, Tony Q. S. Quek, and M. Kountouris. "Cluster-centric Cache Utilization Design in Cooperative Small Cell Networks". In: *IEEE International Conference on Communications (ICC)* (May 2016) (cit. on p. 16).

[Che+18] D. Chen, H. Al-Shatri, T. Mahn, A. Klein, and V. Kuehn. "Energy Efficient Robust F-RAN Downlink Design for Hard and Soft Fronthauling". In: *IEEE 87th Vehicular Technology Conference (VTC Spring)* (Jun 2018) (cit. on pp. 22, 75).

[Che18] D. Chen. "Low Complexity Power Control with Decentralized Fog Computing for Distributed Massive MIMO". In: *IEEE Wireless Communications and Networking Conference (WCNC)* (Apr 2018) (cit. on pp. 22, 165).

[Cis12] Cisco. *[online] http://suo.im/6aGka3*. Tech. rep. 2012 (cit. on p. 77).

[CK16a] D. Chen and V. Kuehn. "Adaptive Radio Unit Selection and Load Balancing in the Downlink of Fog Radio Access Network". In: *IEEE Global Communications Conference (GLOBECOM)* (Dec 2016) (cit. on pp. 22, 75).

[CK16b] D. Chen and V. Kuehn. "Alternating Information Bottleneck Optimization for Weighted Sum Rate and Resource Allocation in the Uplink of C-RAN". In: *20th International ITG Workshop on Smart Antennas (WSA)* (Mar 2016) (cit. on pp. 22, 39).

[CK16c] D. Chen and V. Kuehn. "Optimization Scheme of Noisy Network Coding in the Two Way Relay Channels". In: *IEEE Wireless Communications and Networking Conference (WCNC)* (Apr 2016) (cit. on pp. 22, 39).

[CK16d] D. Chen and V. Kuehn. "Alternating Information Bottleneck Optimization for the Compression in the Uplink of C-RAN". In: *IEEE International Conference on Communications (ICC)* (May 2016) (cit. on pp. 22, 39).

[CK16e] D. Chen and V. Kuehn. "Scalar and Vector Compress and Forward for the Two Way Relay Channels". In: *IEEE 83rd Vehicular Technology Conference (VTC Spring)* (May 2016) (cit. on pp. 22, 39).

[CK16f] D. Chen and V. Kuehn. "Weighted Max-Min Fairness Oriented Load-balancing and Clustering for Multicast Cache-Enabled F-RAN". In: *9th International Symposium on Turbo Codes and Iterative Information Processing (ISTC)* (Sep 2016) (cit. on pp. 22, 75).

[CK17a] D. Chen and V. Kuehn. "Joint Resource Allocation and Power Control for Maximizing the Throughput of Multicast C-RAN". In: *11th International ITG Conference on Systems, Communications and Coding (SCC)* (Feb 2017) (cit. on pp. 22, 75).

[CK17b] D. Chen and V. Kuehn. "An Investigation on Energy and Spectral Efficient Robust Design of Fog Radio Access Network". In: *21th International ITG Workshop on Smart Antennas (WSA)* (Mar 2017) (cit. on pp. 22, 75).

[CK17c] D. Chen and V. Kuehn. "Robust Resource Allocation and Clustering Formulation for Multicast C-RAN with Impaired CSI". In: *IEEE International Conference on Communications (ICC)* (May 2017) (cit. on pp. 22, 75).

[Cla19] R. Clark. *Operators Starting to Face Up to 5G Power Cost*. Report. lightreading, Oct. 2019 (cit. on p. 79).

[CSK16] D. Chen, S. Schedler, and V. Kuehn. "Backhaul Traffic Balancing and Dynamic Content-Centric Clustering for the Downlink of Fog Radio Access Network". In: *IEEE 17th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)* (Jul 2016) (cit. on pp. 22, 75).

[CT91] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 1991 (cit. on pp. 24–26).

[CWB08] E. Candes, M. Wakin, and S. Boyd. "Enhancing Sparsity by Reweighted ℓ_1 Minimization". In: *J Fourier Anal Appl.* 14.5 (Oct 2008) (cit. on pp. 37, 38, 98).

[Dai+15] L. Dai, B. Wang, Y. Yuan, S. Han, C-L I, and Z. Wang. "Non-Orthogonal Multiple Access for 5G: Solutions, Challenges, Opportunities, and Future Research Trends". In: *IEEE Communications Magazine* (Sep 2015) (cit. on p. 3).

[DC15] H. S. Dhillon and G. Caire. "Wireless Backhaul Networks: Capacity Bound, Scalability Analysis and Design Guidelines". In: *IEEE Trans. Wireless Comm.* 14.11 (Nov 2015) (cit. on pp. 6, 57, 83).

[DFO20] M. P. Deisenroth, A. A. Faisal, and C. S. Ong. *Mathematics for Machine Learning*. Cambridge University Press, 2020 (cit. on p. 184).

[DLG16] J. Duan, X. Lagrange, and F. Guilloud. "Performance Analysis of Several Functional Splits in C-RAN". In: *IEEE 83rd Vehicular Technology Conference (VTC Spring)* (May 2016) (cit. on p. 16).

[DM06] K. Derinkuyu and M.C.Pınar. "On the S-procedure and Some Variants". In: *Math.* 64 (Jan 2006) (cit. on p. 38).

[DW16] B. Dai and W.Yu. "Energy Efficiency of Downlink Transmission Strategies for C-RAN". In: *IEEE J. Sel. Areas Commun.* 34 (Apr 2016) (cit. on p. 101).

[DY14] B. Dai and W. Yu. "Sparse Beamforming and User-Centric Clustering for Downlink Cloud Radio Access Network". In: *IEEE Access* (Nov 2014) (cit. on pp. 10, 79).

[DY15] B. Dai and W. Yu. "Backhaul-Aware Multicell Beamforming for Downlink Cloud Radio Access Network". In: *IEEE International Conference on Communications (ICC)* (Jun 2015) (cit. on pp. 10, 79).

[DY16a] B. Dai and W. Yu. "Joint User Association and Content Placement for Cache-enabled Wireless Access Networks". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Mar 2016) (cit. on p. 78).

[DY16b] B. Dai and W. Yu. "Energy Efficiency of Downlink Transmission Strategies for C-RAN". In: *IEEE J. Sel. Areas Commun.* 34.4 (Apr 2016) (cit. on pp. 9, 10, 78, 170).

[Eri16a] Ericsson. *4G/5G RAN architecture: how a split can make the difference*. 2016 (cit. on p. 16).

[Eri16b] Ericsson. *Ericsson Annual Report 2016*. Working Paper. Ericsson, 2016 (cit. on p. 1).

[FKB09] G. Fettweis, M. Krondorf, and S. Bittner. "GFDM - Generalized Frequency Division Multiplexing". In: IEEE 69th Vehicular Technology Conference (VTC Spring). Apr 2009 (cit. on p. 3).

[Fou12] Open Networking Foundation. *Software-Defined Networking: The New Form for Networks*. White Paper. Apr. 2012 (cit. on p. 3).

[Fre09] R. M. Freund. "Introduction to Semidefinite Programming (SDP)". In: *MIT* (2009) (cit. on p. 98).

[GCW12] Q. S. Gong, Z. Y Chen, and G. Wei. "Downlink Multicasting Beamforming with Imperfect CSI on Both Transceiver Sides". In: *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)* (Sep 2012) (cit. on p. 92).

[GK11] A. E. Gamal and Y. H. Kim. *Network Inofrmation Theory*. Cambridge University Press, 2011 (cit. on pp. 3, 176, 178).

[Gol12] A. Goldsmith. *Wireless Communications*. Cambridge University Press, 2012 (cit. on p. 59).

[Gre+15] M. Gregori, J. G. Vilardebò, J. Matamoros, and D. Gündüz. "Joint Transmission and Caching Policy Design for Energy Minimization in the Wireless Backhaul Link". In: *IEEE ISIT* (Jun 2015) (cit. on pp. 16, 83).

[Int12] Intel. *[online] https://software.intel.com/en-us/articles/video-aware-wireless-networks*. Tech. rep. 2012 (cit. on p. 77).

[KPS12] M. Kaliszan, E. Pollakis, and S. Stańczak. "Multigroup Multicast with Application-Layer Coding: Beamforming for Maximum Weighted Sum Rate". In: *IEEE Wireless Communications and Networking Conference (WCNC)* (Apr 2012) (cit. on pp. 138, 141, 142).

[KSL08] E. Karipidis, N. D. Sidiropoulos, and Z. Q. Luo. "Quality of Service and Max-Min Fair Transmit Beamforming to Multiple Cochannel Multicast Groups"". In: *IEEE Signal Processing Magazine* 56.3 (Mar 2008) (cit. on pp. 36, 97, 99, 104, 123, 136, 155).

[LBZ15] L. Liu, S. Bi, and R. Zhang. "Joint Power Control and Fronthaul Rate Allocation for Throughput Maximization in OFDMA-Based Cloud Radio Access Network". In: *IEEE Trans. Wireless Comm.* 63.11 (Nov 2015) (cit. on pp. 10, 170).

[Lim+11] S. H. Lim, Y. H. Kim, A. E. Gamal, and S. Y. Chung. "Noisy Network Coding". In: *IEEE Trans. Inf. Theory* 57.5 (May 2011) (cit. on p. 42).

[Liu+17] J. Liu, B. Bai, J. Zhang, and K. B. Letaief. "Cache Placement in Fog-RANs: From Centralized to Distributed Algorithms". In: *IEEE Trans. Wireless Comm.* 16.11 (Nov 2017) (cit. on p. 78).

[Luo+10] Z. Luo, W. K. Ma, A. M. So, Y. Ye, and S. Zhang. "Semidefinite Relaxation of Quadratic Optimization Problems"". In: *IEEE Signal Processing Magazine* 27.3 (May 2010) (cit. on p. 36).

[LY17] L. Liu and W. Yu. "Cross-Layer Design for Downlink Multi-Hop Cloud Radio Access Networks with Network Coding". In: *IEEE Trans. Signal Processing* 65.7 (Apr 2017) (cit. on p. 10).

[LZ16] L. Liu and R. Zhang. "Downlink SINR balancing in C-RAN under Limited Fronthaul Capacity". In: *IEEE ICASSP 2016* (Mar 2016) (cit. on p. 10).

[Mao+17] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief. "A Survey on Mobile Edge Computing: The Communication Perspective". In: *IEEE Communications Surveys & Tutorials* 19.2322 - 2358 (Aug 2017) (cit. on p. 169).

[Mar+16] T. L. Marzetta, E. G. Larsson, H. Yang, and H. Q. Ngo. *Fundamentals of Massive MIMO*. Cambridge University Press, 2016 (cit. on pp. 3, 16, 17, 166, 168, 175, 177, 182, 183).

[Mar10] T. L. Marzetta. "Noncooperative Cellular Wireless with Unlimited Numbers of Base Station Antennas". In: *IEEE Trans. Wireless Comm.* 9.11 (Nov 2010) (cit. on pp. 3, 16, 17, 166, 175).

[MBe10] M.Bellanger. "FBMC physical layer : a primer". In: *PHYDYAS* (2010) (cit. on p. 3).

[MFR20] W. Melo, M. Fampa, and F. Raupp. "An Overview of MINLP Algorithms and their Implementation in Muriqui Optimizer". In: *Annals of Operations Research* 286.217–241 (Mar 2020) (cit. on pp. 36, 96).

[MN14] M. A. Maddah-Ali and U. Niesen. "Fundamental Limits of Caching". In: *IEEE Trans. Inf. Theory* 60.5 (May 2014) (cit. on pp. 12, 78, 80).

[Mob11] China Mobile. *C-RAN: The road towards green RAN, ver. 2.5*. White Paper. 2011 (cit. on pp. 3, 5).

[Mou+17] G. Mountaser, M. L. Rosas, T. Mahmoodi, and M. Dohler. "On the Feasibility of MAC and PHY Split in Cloud RAN". In: *IEEE Wireless Communications and Networking Conference (WCNC)* (Mar 2017) (cit. on p. 16).

[NCS17] A. Nordrum, K. Clark, and IEEE Spectrum Staff. *5G Bytes: Massive MIMO Explained*. Report. IEEE Spectrum, June 2017 (cit. on p. 3).

[Ne15] Y. Niu and Y. Li *et al. A Survey of Millimeter Wave (mmWave) Communications for 5G: Opportunities and Challenges*. https://arxiv.org/pdf/1502.07228.pdf. 2015 (cit. on pp. 3, 6).

[New20] Abacus News. *5G towers are consuming a lot of energy, so China Unicom is putting some of them to sleep overnight*. Report. Abacus News, Aug. 2020 (cit. on p. 79).

[NN14] S. Nasseri and M. R. Nakhai. "Min-Max Robust Transmit Beamforming for Power Efficient Quality of Service Guarantee". In: *IEEE Global Communications Conference (GLOBECOM)* (Dec 2014) (cit. on p. 92).

[OMM16] A. Osseiran, J. F. Monserrat, and P. Marsch. *5G Mobile and Wireless Communications Technology*. Cambridge University Press, 2016 (cit. on pp. 2, 5, 13).

[Par+13a] S-H Park, O. Simeone, O. Sahin, and S. Shamai (Shitz). "Joint Precoding and Multivariate Fronthaul Compression for the Downlink of Cloud Radio Access Networks". In: *IEEE Trans. Signal Processing* 61.22 (Nov 2013) (cit. on pp. 10, 89, 170).

[Par+13b] S. H. Park, O. Simeone, O. Sahin, and S. Shamai (Shitz). "Joint Decompression and Decoding for Cloud Radio Access Networks"". In: *IEEE Signal Processing Letters* 20.5 (May 2013) (cit. on pp. 5, 10, 42, 170, 171, 183).

[Par+14] S. H. Park, O. Simeone, O. Sahin, and S. Shamai (Shitz). "Fronthaul Compression for Cloud Radio Access Networks". In: *IEEE Signal Processing Magazine* (Nov 2014) (cit. on pp. 5, 7, 9, 10, 18, 45, 46, 58, 170, 176, 183, 184).

[Par+16] S-H Park, O. Simeone, O. Sahin, and S. Shamai (Shitz). "Multihop Backhaul Compression for the Uplink of Cloud Radio Access Networks". In: *IEEE Trans. Vehicular Tech.* 65.5 (May 2016) (cit. on p. 10).

[PCB15] S-H Park, C. Chae, and S. Bahk. "Large-Scale Antenna Operation in Heterogeneous Cloud Radio Access Networks: A Partial Centralization Approach"". In: *IEEE Wireless Communications* 22.3 (Jun 2015) (cit. on p. 20).

[PDY15] P. Patil, B. Dai, and W. Yu. "Performance Comparison of Data-sharing and Compression Strategies for Cloud Radio Access Networks". In: *23rd European Signal Processing Conference (EUSIPCO)* (Aug 2015) (cit. on pp. 10, 78).

[Pen+14] X. Peng, J-C Shen, J. Zhang, and K. B. Letaief. "Joint Data Assignment and Beamforming for Backhaul Limited Caching Networks". In: *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)* (Sep 2014) (cit. on p. 16).

[Pen+15] M. Peng, C. Wang, V. Lau, and H. V. Poor. "Fronthaul-Constrained Cloud Radio Access Networks: Insights and Challenges". In: *IEEE Wireless Communications* 22.2 (Apr 2015) (cit. on pp. 5, 6, 10).

[Pen+16] M. Peng, S. Yan, K. Zhang, and C. Wang. "Fog Computing based Radio Access Networks: Issues and Challenges". In: *IEEE Network* 30.4 (Jul 2016) (cit. on pp. 3, 11, 18, 82, 171).

[PL03] D. Peaucelle and Y. Labit. "User's Guide for SEDUMI INTERFACE". In: *Optimization Online* (Jun 2003) (cit. on p. 98).

[Pon+11] D. Ponukumati, F. F. Gao, M. Bode, and X. W. Liao. "Multicell Downlink Beamforming with Imperfect Channel Knowledge at Both Transceiver Sides". In: *IEEE Comm. Letters* 15.10 (Oct 2011) (cit. on p. 92).

[Pou+16] K. Poularakis, G. Iosifidis, V. Sourlas, and L. Tassiulas. "Exploiting Caching and Multicast for 5G Wireless Networks". In: *IEEE Trans. Wireless Comm.* 15.4 (Apr 2016) (cit. on p. 11).

[PSS16] S-H Park, O. Simeone, and S. Shamai (Shitz). "Joint Optimization of Cloud and Edge Processing for Fog Radio Access Networks". In: *IEEE Trans. Wireless Comm.* 15.11 (Nov 2016) (cit. on pp. 16, 18, 78, 81, 82, 171, 176).

[Que+17] Tony Q. S. Quek, M. Peng, O. Simeone, and W. Yu. *Cloud Radio Access Networks: Principles, Technologies, and Applications*. Cambridge University Press, 2017 (cit. on pp. 5, 10).

[Res20] CVX Research. "CVX: Matlab Software for Disciplined Convex Programming". In: (2020) (cit. on p. 32).

[RHL13] M. Razaviyayn, M. Hong, and Z. Luo. "A Unified Convergence Analysis of Block Successive Minimization Methods for Nonsmooth Optimization". In: *SIAM J. Optim* 23.2 (2013) (cit. on pp. 27, 28, 53).

[Rus+13] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson. "Scaling Up MIMO: Opportunities and Challenges with Very Large Arrays". In: *IEEE Signal Processing Magazine* 30.1 (Jan 2013) (cit. on p. 3).

[SGG18] S. O. Somuyiwa, A. Gyoergy, and D. Guenduez. "A Reinforcement-Learning Approach to Proactive Caching in Wireless Networks". In: *IEEE Journal on Selected Areas in Communications* 36.1331 - 1344 (Jun 2018) (cit. on p. 13).

[Sha+13] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire. "Femto Caching: Wireless Content Delivery Through Distributed Caching Helpers". In: *IEEE Trans. Inf. Theory* 59.12 (Dec 2013) (cit. on pp. 12, 77, 80, 81).

[Sha14] S. Shamai. *On Cloud Radio Access Networks: Information theoretic considerations*. Plenary Talk of ISWCS 2014. Aug. 2014 (cit. on p. 68).

[Sha48] C. Shannon. *A Mathematical Theory of Communication*. WARREN WEAVER, 1948 (cit. on p. 23).

[Son+17] L. Song, R. Wichman, Y. Li, and Z. Han. *Full-Duplex Communications and Networks*. Cambridge University Press, 2017 (cit. on p. 3).

[SYC14] M. Sadeghi, C. Yuen, and Y. H. Chew. "Sum Rate Maximization for Uplink Distributed Massive MIMO Systems with Limited Backhaul Capacity". In: *IEEE Global Communications Conference (GLOBECOM) Workshop* (Dec 2014) (cit. on pp. 20, 170, 171).

[SZL14] Y. M. Shi, J. Zhang, and K. B. Letaief. "Group Sparse Beamforming for Green Cloud-RAN". In: *IEEE Trans. Wireless Comm.* 13.5 (May 2014) (cit. on pp. 5, 10, 79, 82, 170).

[SZL15] Y. M. Shi, J. Zhang, and K. B. Letaief. "Robust Group Sparse Beamforming for Multicast Green Cloud-RAN with Imperfect CSI". In: *IEEE Trans. Signal Processing* 63.17 (Sep 2015) (cit. on p. 92).

[Tao+16] M. Tao, E. Chen, H. Zhou, and W. Yu. "Content-Centric Sparse Multicast Beamforming for Cache-Enabled Cloud RAN". In: *IEEE Trans. Wireless Comm.* 15.9 (Sep 2016) (cit. on pp. 5, 16, 18, 77, 79, 81, 82, 106–115, 124, 126, 129, 171).

[TPB99] N. Tishby, F. C. Pereira, and W. Bialek. "The Information Bottleneck Method". In: *The 37th annual Allerton Conference on Communication, Control, and Computing* (Sep 1999) (cit. on pp. 27–30, 42, 46, 50, 52, 54).

[TT11] K-C. Toh and M. J. Todd. "On the Implementation and Usage of SDPT3 – A Matlab Software Package for Semidefinite-Quadratic-Linear Programming". In: *Handbook on Semidefinite, Conic and Polynomial Optimization* 715-754 (Sep 2011) (cit. on p. 98).

[TTJ15] O. Tervo, L. Tran, and M. Juntti. "Optimal Energy-Efficient Transmit Beamforming for Multi-User MISO Downlink". In: *IEEE Trans. Signal Processing* 63.20 (Oct 2015) (cit. on p. 82).

[UAS16] Y. Ugur, Z. H. Awan, and A. Sezgin. "Cloud Radio Access Networks with Coded Caching". In: *ITG Workshop on Smart Antennas* (Mar 2016) (cit. on pp. 16, 79, 82).

[Wan+14] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. Leung. "Cache in the Air: Exploiting Content Caching and Delivery Techniques for 5G Systems". In: *IEEE Communications Magazine* (Feb 2014) (cit. on pp. 12, 80).

[Win14] A. Winkelbauer. "Dissertation: Blind Performance Estimation and Quantizer Design with Applications to Relay Networks". In: (Dec 2014) (cit. on p. 31).

[Won+17] V. Wong, R. Schober, D. W. K. Ng, and L. Wang. *Key Technologies for 5G Wireless Systems*. Cambridge University Press, 2017 (cit. on pp. 3, 5).

[Wue+14] D. Wuebben, P. Rost, J. S. Bartelt, M. Lalam, V. Savin, M. Gorgoglione, A. Dekorsy, and G. Fettweis. "Benefits and Impact of Cloud Computing on 5G Signal Processing". In: *IEEE Signal Processing Magazine* (Nov 2014) (cit. on pp. 5, 13, 16).

[WZ76] A. D. Wyner and Jacob Ziv. "The Rate-Distortion Function for Source Coding with Side Information at the Decoder". In: *IEEE Trans. Inf. Theory* 22.1 (Jan 1976) (cit. on pp. 9, 27, 42).

[Zei11] G. C. Zeitler. "Low-Precision Quantizer Design for Communication Problems". In: *http://mediatum.ub.tum.de/doc/1092069/1092069.pdf* (Nov 2011) (cit. on pp. 31, 48, 50, 54, 62).

[ZY14] Y. H. Zhou and W. Yu. "Optimized Backhaul Compression for Uplink Cloud Radio Access Network". In: *IEEE J. Sel. Areas Commun.* 32.6 (Jun 2014) (cit. on pp. 5, 9, 42, 68, 170, 176).

Curriculum Vitae

Personal Information

Last name:	Chen
First name:	Di
Date of Birth:	28.01.1988
Place of Birth:	Shandong, China
E-mail:	chendi880128@gmail.com

Education

03.2010 - 10.2012	Master study at University of Ulm, Ulm, Germany
02.2008 - 07.2008	Bachelor exchange program at Sun Yat-sen University, Zhuhai, China
09.2006 - 03.2010	Bachelor study at Shandong University, Jinan, China
09.2003 - 06.2006	No. 1 High School, Dezhou, China
09.2000 - 06.2003	No. 9 Middle School, Dezhou, China
09.1994 - 06.2000	Tianqudonglu Primary School, Dezhou, China

Career

10.2017 - present	Software Developer at Nokia, Ulm, Germany
07.2014 - 09.2017	Research Fellow at Institute of Communications Engineering, University of Rostock, Rostock, Germany
11.2012 - 06.2014	Research Fellow at Chair of Digital Communication Systems, Ruhr University Bochum, Bochum, Germany
04.2012 - 07.2012	Research Student at Daimler AG, Ulm, Germany
08.2011 - 09.2012	Research Student at Institute of Communications Engineering, University of Ulm, Ulm, Germany

Lebenslauf

Persönliche Daten

Name:	Chen
Vorname:	Di
Geburtsdatum:	28.01.1988
Geburtsort:	Shandong, China
E-mail:	chendi880128@gmail.com

Bildungserfahrung

03.2010 - 10.2012	Master-Studium an der Universität Ulm, Ulm, Deutschland
02.2008 - 07.2008	Bachelor-Austauschprogramm an der Sun-Yat-sen-Universität, Zhuhai, China
09.2006 - 03.2010	Bachelor-Studium an der Shandong-Universität, Jinan, China
09.2003 - 06.2006	Oberschule No. 1, Dezhou, China
09.2000 - 06.2003	Mittelschule No. 9, Dezhou, China
09.1994 - 06.2000	Grundschule Tianqudonglu, Dezhou, China

Berufserfahrung

10.2017 - heute	Softwareentwickler bei Nokia, Ulm, Deutschland
07.2014 - 09.2017	Wissenschaftlicher Mitarbeiter am Institut für Nachrichtentechnik der Universität Rostock, Rostock, Deutschland
11.2012 - 06.2014	Wissenschaftlicher Mitarbeiter am Lehrstuhl für Digitale Kommunikationssysteme der Ruhr-Universität Bochum, Bochum, Deutschland
04.2012 - 07.2012	Studentische Hilfskraft bei Daimler AG, Ulm, Deutschland
08.2011 - 09.2012	Studentische Hilfskraft am Institut für Nachrichtentechnik der Universität Ulm, Ulm, Deutschland

Authenticity Declaration / Selbständigkeitserklärung

I hereby declare that I wrote this dissertation independently and I did not use any unnamed sources or aid.

I have clearly referenced all sources used in the work.

This dissertation has not been published and was neither in full, nor in similar form, previously submitted for grading at any academic institution.

Signature:

Ulm, on September 13, 2021, Di Chen

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Alle Stellen, die wörtlich oder sinngemäß aus Veröffentlichungen entnommen sind, sind als solche kenntlich gemacht.

Die Arbeit ist noch nicht veröffentlicht und ist in ähnlicher oder gleicher Weise noch nicht als Prüfungsleistung zur Anerkennung oder Bewertung vorgelegt worden.

Unterschrift:

Ulm, den September 13, 2021, Di Chen