Beiträge aus der Informationstechnik

Mobile Nachrichtenübertragung
Nr. 84

Jens Steven Bartelt

Joint Design of Access and Fronthaul Uplinks in Cloud Radio Access Networks

VOGT

Dresden 2017

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.dnb.de abrufbar.

Bibliographic Information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at http://dnb.dnb.de.

Zugl.: Dresden, Techn. Univ., Diss., 2017

Die vorliegende Arbeit stimmt mit dem Original der Dissertation „Joint Design of Access and Fronthaul Uplinks in Cloud Radio Access Networks“ von Jens Steven Bartelt überein.

© Jörg Vogt Verlag 2017
Alle Rechte vorbehalten. All rights reserved.

Gesetzt vom Autor

ISBN 978-3-95947-011-7

Jörg Vogt Verlag
Niederwaldstr. 36
01277 Dresden
Germany

Phone: +49-(0)351-31403921
Telefax: +49-(0)351-31403918
e-mail: info@vogtverlag.de
Internet : www.vogtverlag.de

Technische Universität Dresden

Joint Design of Access and Fronthaul Uplinks in Cloud Radio Access Networks

Jens Steven Bartelt

von der Fakultät Elektrotechnik und Informationstechnik der
Technischen Universität Dresden
zur Erlangung des akademischen Grades eines

Doktoringenieurs
(Dr.-Ing.)

genehmigte Dissertation

Vorsitzender:	Prof. Dr. sc. techn. habil. Frank Ellinger
Gutachter:	Prof. Dr.-Ing. Dr. h.c. Gerhard Fettweis
	Prof. Dr.-Ing. Volker Kühn
Tag der Einreichung:	24.11.2016
Tag der Verteidigung:	09.03.2017

Jens Steven Bartelt
Joint Design of Access and Fronthaul Uplinks in Cloud Radio Access Networks
Dissertation, 09. März 2017
Vorsitzender: Prof. Dr. sc. techn. habil. Frank Ellinger
Gutachter: Prof. Dr.-Ing. Dr. h.c. Gerhard Fettweis und Prof. Dr.-Ing. Volker Kühn

Technische Universität Dresden
Vodafone Chair für Mobile Nachrichtensysteme (MNS)
Institut für Nachrichtentechnik
Fakultät Elektrotechnik und Informationstechnik
01062 Dresden

If I had asked people what they wanted, they would have said faster horses.

— **Henry Ford**
(Founder of the Ford Motor Company)

Abstract

Cloud radio access networks (C-RANs) are a promising concept for the architecture of 5th generation mobile networks. In C-RANs, signal processing is not performed at the access points as in common mobile networks, but is instead centralized in large, cloud-based data centers. This approach promises many benefits, including smaller-footprint base stations, simplified network management, maintenance and upgrades, economies of scale and a more efficient implementation of cooperative processing techniques. However, such a centralized architecture comes at the heavy price of an extensive, so-called fronthaul network, which has to exchange the raw, unprocessed radio signals between the remote access points and the central processing unit. This requires the fronthaul network to fulfill challenging requirements in terms of data rate, latency, and synchronization. Currently, these fronthaul networks are designed, deployed, and operated separately from the radio access network, meaning that there is little cooperation and information exchange between the fronthaul and radio access network. To mitigate this, this thesis proposes a *joint* design of the radio access and fronthaul links, by considering the impact that they have on one another, and by exchanging more side-information to form a joint radio access/fronthaul link.

A first step towards such a joint design is the re-design of the so-called functional split, which refers to the amount of processing performed at the remote access points versus that performed at the central unit. While the current approaches are two extreme cases of either full centralization or decentralization, an intermediate option can reduce the strict fronthaul requirements, while maintaining several of the benefits of centralization.

Furthermore, the access and fronthaul links can cooperate on the level of the physical interface by exchanging information about signal statistics and their respective channel qualities. With this, a joint minimum mean square error receiver is designed in this thesis, aiming to improve the performance of the joint radio access and fronthaul link. This receiver is especially beneficial when

utilizing wireless millimeter wave fronthaul links, which suffer from a reduced channel quality due to their high attenuation by precipitation.

Finally, it can be shown that the design of the quantizer, which is employed to digitize the radio signals before fronthaul transmission, has a considerable impact on overall performance. Accordingly, optimization schemes are proposed to limit this impact by optimized quantizer design.

The three proposed approaches – intermediate functional split, joint receivers, and optimized quantizer design – are shown to improve the end-to-end performance of the joint radio access/fronthaul link considerably in a relevant scenario. An analysis of their implementation complexity further underlines their feasibility. In summary, the joint design of radio access and fronthaul is a promising novel approach to solve the challenges of today's fronthaul networks.

Kurzfassung

Cloud-basierte Mobilfunknetzte (C-RANs) sind ein vielversprechender Ansatz für die Architektur der Mobilfunknetzte der fünften Generation. In C-RANs wird Signalverarbeitung nicht nur an den Zugangspunkten ausgeführt, wie es in konventionellen Architekturen der Fall ist, sondern auch in großen, cloud-basierten Datenzentren. Dieser Ansatz verspricht viele Vorteile: kleinere und leichtere Basisstationen, vereinfachte Verwaltung, Wartung und Nachrüstung, Rationalisierungseffekte, sowie eine effizientere Implementierung von kooperativen Signalverarbeitungstechniken. Allerdings benötigt eine solch zentralisierte Architektur ein aufwendiges, sogenanntes Fronthaul-Netzwerk, welches, rohe, unverarbeitete Funksignale zwischen den verteilt aufgestellten Zugangspunkten und der zentralisierten Signalverarbeitungsprozessoren austauscht. Dazu muss dieses Fronthaul-Netzwerk herausfordernde Anforderungen bezüglich Datenrate, Latenz und Synchronisierung erfüllen. Zurzeit werden die Fronthaul-Netzwerke getrennt von den eigentlichen Zugangsnetzwerken entwickelt, aufgebaut und betrieben. Die hat zur Folge, dass sehr wenig Informationen zwischen beiden Netzsegmenten ausgetauscht werden und sie kaum kooperativ arbeiten. Um dies zu verbessern, wird in dieser Arbeit vorgeschlagen, Fronthaul- und Zugangsnetzte gemeinsam zu betrachten, indem ihre gegenseitige Abhängigkeit besser berücksichtigt wird und mehr Informationen zwischen ihnen ausgetauscht werden, um so eine gemeinsame Funkzugangs-und Fronthaul-Strecke zu bilden.

Ein erster Schritt in dieser Richtung, ist der Neuentwurf der Aufgabenteilung bei der Signalverteilung zwischen den Zugangspunkten und der zentralen Cloud. Die heutige Aufgabenverteilung stellt dabei zwei Extreme dar, bei der entweder sämtliche Signalverarbeitung in der Basisstation am Zugangspunkt stattfindet, oder sie komplett in die Cloud ausgelagert wird. Ein dazwischenliegender Ansatz kann stattdessen die Anforderungen an das Fronthaul-Netzwerk senken und zur selben Zeit einige Vorteile der Zentralisierung erhalten.

Des Weiteren können Funkzugangs- und Fronthaul-Verbindung auch auf der physikalischen Netzwerkschicht kooperieren, indem Informationen über Statis-

tik und Qualität ihrer Signale und Kanäle ausgetauscht werden. Mit diesem Ansatz wird in dieser Arbeit ein kooperativer Empfänger zur Minimierung des quadratischen Fehlers entworfen, welcher die Übertragungsqualität der gemeinsamen Funkzugangs- und Fronthaul-Verbindung erhöht. Dieser Empfänger ist insbesondere von Vorteil, wenn Fronthaul-Verbindungen mit drahtloser Millimeterwellentechnik realisiert werden, deren Übertragungsqualität durch einen hohen Pfadverlust und Dämpfungseffekte bei Niederschlag limitiert ist.

Weiterhin wird gezeigt, dass das Design des Quantisierers, der notwendig ist um die analogen Funksignale für die Fronthaul-Übertragung zu digitalisieren, einen entscheidenden Einfluss auf die Qualität der Übertragung hat. Daher werden in dieser Arbeit Optimierungsverfahren entwickelt, um diesen Einfluss durch ein angepasstes Quantisiererdesign zu limitieren. Die drei vorgeschlagenen Ansätze – Aufgabenverteilung zwischen Zugangspunkt und Cloud, kooperativer Empfänger und optimierter Quantiserer – zeigen in einem repräsentativen Szenario eine deutliche Verbesserung der Ende-zu-Ende Übertragung der gemeinsamen Zugangs-/Fronthaul-Verbindung. Des Weiteren unterstreicht eine Analyse der Implementierungskomplexität die Praktikabilität der Ansätze.

Zusammenfassend betrachtet ist der gemeinsame Entwurf von Zugangs- und Fronthaul-Netzwerk ein neuer, vielversprechender Ansatz um den Herausforderungen zukünftiger Mobilfunknetzte zu begegnen.

Acknowledgement

This thesis is the combined output of four years of hard work. Fortunately, I was not alone in those four years, or otherwise this work would never have been possible. Accordingly, some thanks are in order.

First and foremost, I would like to thank Prof. Gerhard Fettweis, who has been my stout advisor and the tireless leader of the Vodafone Chair during this whole time. I think he does a fantastic job in providing an environment for PhD students to flourish, research to advance, and the future to being shaped.

I would also like to thank Prof. Volker Kühn, for taking time to review my thesis and providing valuable feedback.

Next, I would like to thank Dan Zhang, who was my research group leader for the latter part of my time at the chair, and provided me with valuable input, not only in matters scientific but also in the complicated jungle of academic procedures.

I had a great time with my fellow PhD students as well, and had the pleasure to go to awesome trips with many of them. As the joyful experiences I had are too many to name, I would like to exemplary thank: Björn Almeroth, who introduced me to the pleasures of being a teaching assistant; Lukas Landau, who made us win more awards than we deserved; Hermann Hensel, who implemented what I could not; David Öhmann, who pushed me by his example; Mario Schiselski, who supported my secondary career as actor and director; and finally Steffen Watzek, Kathrin Fromke and Sylvia Steppat, for keeping the chair running and supporting me and everyone in all matters organizational.

During my time at the chair, I had the pleasure of working in two large EU projects, iJOIN and 5G-XHaul. While the inevitable organizational overhead of these projects sometimes seemed to overbear the scientific work, I find that the projects were a main reason for me to stand where I stand today. I had a great time with the colleagues in these projects and they provided input to my research where others could not. Hence, I would like to specifically thank Peter Rost and Dirk Wübben from iJOIN, and Eckhard Grass, Jesús Gutiérrez and Daniel Camps-Mur from 5G-XHaul, for their guidance, suggestions, and leadership.

Fortunately, life is not only about work, so I thank Philip Frank, David Teich and Sebastian Häfner for their company during uncounted hours of lunch and coffee.

My studies would not have been possible without my family, who supported me since I moved to Dresden nine years ago, first financially, then morally. Thank you for being there and sorry I did not manage to call more often.

Finally, my heartfelt thanks to the love of my life, Thekla Pufe. You make my day, every day.

Contents

List of Figures xvii

List of Tables xix

List of Symbols xxi

List of Abbreviations xxvii

1 Introduction 1

1.1 Motivation 1

1.2 Contribution and Outline 2

2 Centralization Aspects of Mobile Networks 5

2.1 Cloud Radio Access Network Architectures 5

2.1.1 From Distributed to Centralized Radio Access Networks . . 5

2.1.2 The Road to Cloud Radio Access Networks 7

2.1.3 Functional Splits 9

2.2 Requirements of Future Transport Networks 14

2.2.1 5G Radio Access Technologies 14

2.2.2 Data Rate 18

2.2.3 Latency 26

2.2.4 Time Synchronization 29

2.2.5 Frequency Synchronization 31

2.2.6 Reliability 32

2.3 Enablers for Future Transport Networks 32

2.4 Chapter Summary 35

3 Joint Design of Radio Access and Fronthaul 37

3.1 System Model 37

3.1.1 Model Overview 37

3.1.2 Transmitter and Receiver 38

3.1.3 Quantizer 44

3.1.4 Radio Access Channel 47
3.1.5 Fronthaul Channel 49
3.1.6 System Model Summary 50
3.2 Properties of the Joint Access and Fronthaul Channel 52
3.3 Joint Design Concept 56
3.4 Chapter Summary 58

4 MMSE Receivers for Fronthaul Networks 59
4.1 General Variant 59
4.2 Simplified Variant 62
4.3 Functional Split Variants 63
4.4 Iterative Variant 65
4.5 Chapter Summary 68

5 Optimized Quantization 71
5.1 Quantization Impact 71
5.2 Optimized Decision Thresholds 73
5.3 Optimized Power Allocation 78
5.4 Chapter Summary 82

6 Performance Analysis 85
6.1 Overview of Different Approaches 85
6.2 Use Case and Scenario 87
6.3 Results 89
6.3.1 Impact on Different Signal Types 89
6.3.2 Impact of the Functional Split 94
6.3.3 Impact of the Fronthaul Channel Model 95
6.3.4 Impact of the Quantizer Optimization 96
6.4 Chapter Summary 98

7 Implementation Aspects 101
7.1 Evaluation Metrics 101
7.2 Conventional Receiver vs. Joint Receiver 102
7.2.1 Computational Complexity and Latency 102
7.2.2 Hardware Architecture 103
7.2.3 Pre-calculated Look-up Tables 105
7.3 Implementation of Optimization Schemes 106
7.4 Network Architecture 107
7.5 Chapter Summary 110

8 Conclusions **111**
8.1 Summary 111
8.2 Core Findings 111
8.3 Future Work 113

A Appendix **115**
A.1 Derivation of Noise Contributions 115
A.2 Derivative of the Mutual Information 118
A.3 Relation Between Total Noise, Quantization Noise and FH Noise . 119
A.4 Pathloss and Range of a Millimeter Wave Fronthaul Link 120
A.5 Simulation Parameters 121

Bibliography **123**

Publications and Patents **133**

Curriculum Vitae **135**

List of Figures

2.1 Comparison of D-RAN, centralized RAN (center), and Cloud-RAN . 7
2.2 Signal processing chain and functional split options. 10
2.3 Simplified LTE resource grid 20
2.4 Fronthaul throughput requirements 21
2.5 Aggregated data rate PDFs 23
2.6 Measured MCS distribution and resource utilization 24
2.7 Busy hour and overall load CCDF 25
2.8 FH data rate CCDF 25
2.9 Required aggregated FH capacity 26
2.10 FH round-trip latencies 30

3.1 General system model. 38
3.2 System model for user data transmitter. 39
3.3 System model for user data receiver. 39
3.4 Synchronization signal transmitter. 42
3.5 Timing advance estimation receiver. 42
3.6 System model of transmitter for reference signals. 44
3.7 System model for channel estimation. 44
3.8 Illustration of quantization notation. 45
3.9 System model for radio access channel. 48
3.10 System model for the binary symmetric FH channel. 50
3.11 Summary of different system model building blocks. 51
3.12 Different noise contributions to joint RAN/FH channel 53
3.13 Conditional probabilities of received signal amplitude 56
3.14 MI of the joint RAN/FH channel 56
3.15 Illustration of the concept of joint RAN/FH design. 57

4.1 System model for the MMSE receiver. 60
4.2 MSE and MI for the MMSE receiver 61
4.3 Performance of the simplified variant compared to the general variant 63
4.4 PDFs of different signals for amplitude and probability split 64
4.5 PDFs of different signals for soft-bit split 64

4.6 Iterative extension of the MMSE receiver 66
4.7 Illustration of the impact of the iterative scheme 66
4.8 BER of iterative vs. non-iterative MMSE receiver 67
4.9 Illustration of impact of iterative scheme 67

5.1 Impact of quantizer resolution B on different noise contributions . 72
5.2 MI, remote noise, and total noise for different dynamic ranges . . . 75
5.3 MI for large dynamic ranges for $\mathrm{SNR}_{\mathrm{FH}} = 3.3$ dB. 75
5.4 Optimized dynamic range 75
5.5 Results of threshold optimization 77
5.6 Optimized power allocation. 78
5.7 Results of optimized power allocation. 82

6.1 Compatibility of different approaches 85
6.2 $\mathrm{SNR}_{\mathrm{FH}}$ versus link distance for different rain rates 88
6.3 BERs of different scenarios and receivers for different RAN MCS. . 90
6.4 Throughput of different scenarios and receivers with uncoded FH. . 92
6.5 Throughput comparison of MMSE receiver and conventional receiver 92
6.6 Throughput evaluation. 93
6.7 Normalized channel estimation error 94
6.8 Missed detection probability 94
6.9 BER of different functional splits using the MMSE receiver. 95
6.10 BER using AWGN FH channel and BSC FH channel 96
6.11 Impact of dynamic range optimization on BER 97
6.12 Comparison of performance 98

7.1 Comparison of receiver structures 102
7.2 Processing architecture of MMSE receiver. 104
7.3 Detail of processing architecture 104
7.4 Detail of optimized architecture 105
7.5 Computational complexity 105
7.6 System architecture for different centralization schemes. 108
7.7 Data rate for different cloud deployments 108

List of Tables

2.1 Exemplary parametrization and requirements of 5G RATs. 17
2.2 Parameters for FH latency estimation. 30

3.1 Quantization and quantization demapping example 45

6.1 Overview of different approaches. 86
6.2 Parameters for exemplary 60 GHz FH link. 88
6.3 Rain attenuation at 60 GHz based on [ITU12]. 89

7.1 Complexity of the MMSE receiver. 104

A.1 Simulation parameters. 121

List of Symbols

Notation

$\mathbf{x}$	Vector
x	Scalar, element of vector $\mathbf{x}$
$\hat{x}$	Estimate of x
x^*	Optimal value of x
$\overline{x}$	Average value of x
x_q	Quantized version of x
X	Fourier transform of $\mathbf{x}$
$\mathbf{X}$	Matrix
$\mathcal{X}$	Set
$\lvert\mathcal{X}\rvert$	Cardinality of set $\mathcal{X}$
$f(\cdot)$	Mathematical function
$F(\cdot)$	Processing function, system

Symbols

a_1, a_2	Zadoff-Chu indices
B	Number of bits
BER	Bit-error rate
BLER	Block-error rate
b	Bit index
BW	Channel bandwidth
c	Quantizer input
c'	Demapped quantizer output
c_q	Quantizer codebook entry
$c_{q'}$	Quantizer codebook entry after fronthaul
D	Data rate

D_{TP}	Throughput
d	Uncoded, transmit data bit
d'	Received, decoded data bit
$d_{\text{RU,CU}}$	Distance between remote and central unit
$d_{\text{TX,RX}}$	Distance between transmitter and receiver
e	Euler's number
F	Error between quantized symbol and received quantized symbol
F_{RX}	Receiver noise figure
f_{C}	Carrier frequency
f_{S}	Sampling rate
$f_{\text{S,PRACH}}$	Sampling rate of the physical random access channel
$G_{\text{A,TX}}$	Transmit antenna gain
$G_{\text{A,RX}}$	Transmit antenna gain
$G_{\text{a,RX}}$	Analog losses
$G_{\text{d,RX}}$	Digital losses
G_{ox}	Oxygen attenuation
G_{rain}	Rain attenuation
H	Channel frequency respnonse
h	Channel coefficient
I	Identity matrix
$\mathcal{I}$	Set of all quantizer codebook indices
i_b	Codebook index
i_b	b-th bit of codebook index i
k	Zadoff-Chu shifting index
L_{TX}	Transmit power level
L_{EIRP}	Effective isotropic radiated power level
$L_{n,\text{RX}}$	Receiver noise power level
L_{RX}	Received power level
L_{ref}	Reference pathloss
l	Log-likelihood value
m_{FH}	Fronthaul modulation order
$\mathcal{M}_{\text{RAN}}$	Radio access modulation symbol set
m_{RAN}	RAN modulation order
N_{A}	Number of antennas
N_{block}	Block length
N_{FFT}	Fast Fourier transform size
$N_{\text{it,dec}}$	Number of turbo decoder iterations
$N_{\text{it,est}}$	Number of estimation iterations

$N_{\text{it,tot}}$	Total number of iterations
N_{L}	Number of spatial layers
N_{node}	Number of transport nodes
N_{P}	Number of antenna ports/ ADC/DAC chains
N_{Q}	Number of quantization bits
N_{SC}	Number of subcarriers
N_{Sy}	Number of symbols per frame
N_{t}	Noise power density
N_{ZC}	Length of Zadoff-Chu sequence
n_{FH}	Fronthaul noise sample
n_{RAN}	Radio access noise sample
P_e	Error probability
p_{miss}	Missed-detection probability
Q	Matrix of binary quantization indices
q	Quantization index
q_b	b-th bit of binary quantization index q
q'	Received quantization index after fronthaul
$R^{\text{C}}_{\text{RAN}}$	Code rate of the radio access
SNR_{FH}	Fronthaul signal to noise ratio
SNR_{RAN}	Radio access signal to noise ratio
SQR	Signal to quantization noise ratio
$\mathcal{T}$	Set of all quantizer thresholds
T_{c}	Channel coherence time
T_{S}	Subframe duration
T_{FH}	Fronthaul delay
T_{proc}	Processing delay
T_{prop}	propagation delay
T_{RAN}	RAN delay
t_q	Quantizer threshold
u	Encoded data bit
v_{light}	Speed of light
v_{UE}	User speed
x	Radio access transmit symbol
x'	Radio access received symbol
x''	Radio access received symbol after fronthaul and quantizer demapping
x_{corr}	Correlation sequence
x_{TA}	Timing advance correlation peak
y	Fronthaul transmit symbol

y'	Fronthaul received symbol
z	Result of remote processing
z'	Received result of remote processing after fronthaul
z_q	Quantized result of remote processing
$\mathbb{Z}$	Set of integer numbers
α	Scaling parameter
Δc	Quantizer dynamic range
Δt	Quantizer step height
Δf_{S}	Syntonization accuracy
Δk_{TA}	Timing advance index offset
ΔT_{lat}	Delay accuracy
ΔT_{TA}	Timing advance
γ	Scaled signal to noise ratio
η	Reference overhead
γ_b	Scaled SNR per bit
$\overline{\gamma}$	Average scaled SNR per bit
λ	Soft-bit
λ_q	Quantization index of soft-bit
λ_b	Lagrange multiplier
μ	Load, utilization
ν	Lagrange multiplier
σ^2_{BSC}	Fronthaul binary symmetric channel noise variance
σ^2_{Ch}	Channel estimation error
σ^2_{FH}	Fronthaul additive white Gaussian noisee variance
σ^2_{h}	Rayleigh distribution variance
σ^2_{Q}	Quantization noise variance
σ^2_{R}	Remote noise variance
σ^2_{Ray}	Rayleigh channel transition variance
σ^2_{RAN}	Radio access noise variance
σ^2_{tot}	Total noise of joint radio access/ fronthaul channel
ζ	Transport overhead

Functions

$\arg\max(\cdot)$	Argument which returns the maximum value

$\arg\min(\cdot)$	Argument which returns the minimum value
$C(\cdot)$	Encoding
$C^{-1}(\cdot)$	Decoding
$\mathrm{diag}(\cdot)$	Diagonal of a matrix
$\mathrm{E}[\cdot]$	Expected value
$\mathrm{erfc}(\cdot)$	Complimentary error function
$\exp(\cdot)$	Exponential function with base e
$\mathfrak{F}(\cdot)$	Fourier transformation
$\mathfrak{F}^{-1}(\cdot)$	Inverse Fourier transformation
$\gcd(\cdot,\cdot)$	Greatest common divider
$H(\cdot)$	Channel distortion
$H^{-1}(\cdot)$	Equalization
$I(\cdot;\cdot)$	Mutual information
$\mathfrak{Im}(\cdot)$	Imaginary part of a complex number
$log(\cdot)$	Logarithm
$M(\cdot)$	Modulation
$M^{-1}(\cdot)$	Demodulation
$\max(\cdot)$	Maximum value
$\min(\cdot)$	Minimum value
$p(\cdot)$	Probability
$Q(\cdot)$	Quantization
$Q^{-1}(\cdot)$	Quantizer demapping
$\mathcal{Q}(\cdot)$	Q-function
$\mathfrak{Re}(\cdot)$	Real part of a complex number
$R_{\mathrm{C},1}$	Parallel central unit processing
$R_{\mathrm{C},2}$	Serial central unit processing
R_{R}	Remote unit processing
$T(\cdot)$	Transmit processing
$\mathcal{W}(\cdot)$	lambert-W function
$\lvert\cdot\rvert_0$	0-norm, number of non-zero elements
^{H}	Hermitian transpose
$*$	Convolution

List of Abbreviations

3G/4G/5G	3rd/4th/5th Generation of Mobile Networks
ACK	Acknowledgment
ACM	Adaptive Coding and Modulation
A/D	Analog-to-Digital
ADC	Analog-to-Digital Converter
ASIC	Application Specific Integrated Circuit
AWGN	Additive White Gaussian Noise
BBU	Baseband Unit
BER	Bit-Error Rate
BH	Backhaul
BLER	Block-error Rate
BPSK	Binary Phase Shift Keying
BS	Base Station
BSC	Binary Symmetric Channel
CA	Carrier Aggregation
CAPEX	Capital Expenditure
CCDF	Complementary Cumulative Distribution Function
CDF	Cumulative Distribution Function
CFR	Channel Frequency Response
CIR	Channel Impulse Response
CoMP	Coordinated Multi-Point
CPRI	Common Public Radio interface
CQI	Channel Quality Information/Indicator
C-RAN	Cloud Radio Access Network
CRC	Cyclic Redundancy Check
CSI	Channel State Information
CU	Central Unit
DAC	Digital-to-Analog Converter
DAS	Distributed Antenna System

DL	Downlink
D-RAN	Distributed Radio Access Network
DSL	Digital Subscriber Line
DSP	Digital Signal Processor
EIRP	Effective isotropic radiated power
FDE	Frequency Domain Equalization
FEC	Forward Error Correction
FFT	Fast Fourier Transformation
FH	Fronthaul
FPGA	Field Programmable Gate Array
FSO	Free-Space Optics
Gbps	Gigabits Per Second
GPP	General Purpose Processor
HARQ	Hybrid Automatic Repeat Request
I/Q	In-phase/Quadrature-phase
IT	Information Technology
LDPC	Low-Density Parity-Check
LLR	Log-Likelihood Ratio
LOS	Line-Of-Sight
LTE	Long Term Evolution/the 4th Generation of Mobile Networks
MAC	Medium Access Control
Mbps	Megabits Per Second
MCS	Modulation and Coding Scheme
MI	Mutual Information
MIMO	Multiple-Input/Multiple-Output
mmWave	Millimeter Wave
MMSE	Minimum Mean-Square Estimator
MSE	Mean Square Error
MTC	Machine Type Communication
NACK	Negative Acknowledgment
NFV	network Function Virtualization
NLOS	Non-Line-Of-Sight
OFDM	orthogonal Frequency Divison Multiplexing
OPEX	Operational Expenditure
PAPR	Peak-to-Average-Power-Ratio
PDCP	Packet Data Convergence Protocol
PDF	Probability Density Function
PDP	Power Delay Profile

PHY	Physical Layer
PON	Passive Optical Network
ppb	Parts Per Billion
PRACH	Physical Random Access Channel
PTP	Precision Time Protocol
QAM	Quadrature Amplitude Modulation
QoS	Quality of Service
QoE	Quality of Experience
RAN	Radio Access Network/Link
RAT	Radio Access Technology
RLC	Radio Link Control
RRH	Remote Radio Head
RRM	Radio Resource Management
RU	Remote Unit
SC-FDMA	Single Carrier Frequency Division Multiple Access
SDN	Software-Defined Networking
SNR	Signal-to-Noise Ratio
SON	Self-Organizing Network
SQR	Signal-to-Quantization-noise Ratio
SyncE	Synchronous Ethernet
TA	Timing Advance
TC	Traffic Class
TCAM	Ternary Content Addressable Memory
TP	Throughput
TSON	Time-Shared Optical Network
UL	Uplink
WDM	Wavelength Division Multiplexing
ZC	Zadoff-Chu
ZF	Zero Forcing

Introduction

1

1.1 Motivation

For over four decades, cellular mobile communication has been a major success story and is one of the key technologies of today's information society. With the introduction of mobile broadband in the 3rd (3G) and 4th generation (4G) of cellular networks, high-speed internet on-the-go has become an integral part of the life of billions of people. And people want more: according to Ericsson's annual study, the number of broadband subscriptions has risen 20 percent year after year. Between the years 2015 and 2016, mobile data traffic has increased by 60 percent, accumulating to more than 5.5 Zetabytes (5.5 billion Terabyte) per month [Eri16]. The telecommunications industry has already agreed that current 4G networks will not suffice to sate the ever increasing hunger for more data. Hence, global initiatives, such has the NGMN [NGMb], the European Union's 5GPPP [5GP], and 3GPP [3GP] are already working on the 5th generation (5G) of mobile networks, to be ready for deployment around the year 2020.

However, while mobile traffic demand continues to rise, the revenue of mobile operators, who are responsible for deploying and managing the infrastructure, have not kept pace with the increasing cost for ever more advanced and numerous equipment [FF+11]. A more cost-efficient network architecture has hence been a declared target of many stakeholders for the last few years. One approach for this, borrowed from the information technology (IT) industry, are so-called cloud radio access networks (C-RANs) [IHD+14; CCY+15; SRS+13; RBD+14]. By centralizing baseband processing equipment in large data centers or small server rooms, economics of scale can be exploited: management, updates, and maintenance can be largely performed at only a few geographical locations, equipment can be standardized, and hardware can be shared among many tasks thereby avoiding under-utilization and over-provisioning of processors.

However, as the actual radio antennas still need to be deployed remotely, these C-RANs come at the heavy price of an extensive, so-called fronthaul (FH) network, which connects the remote radio access points with the centralized cloud [CPC+13; PCSD15; PWLP15]. This FH network demands the deployment of very high data rate fiber links, and thereby, threatens to nullify the so urgently hoped-for cost saving of centralized networks [Rea15]. Currently, such centralized RAN approaches are already being utilized in 4G networks, however, using a straight-forward, almost brute-force way of implementation, which is one of the reasons for the need of such a cost-intensive FH infrastructure. With 5G networks adding even more demands to the architecture, the FH challenges can only be expected to increase in future networks. To reduce the cost of FH networks, high-capacity wireless FH links have been shown to have great potential by replacing the currently predominant fiber links on the last mile [Eri15; Wel09; AMC15; CHL+15; CHLG16]. However, wireless FH links are less reliable due to their higher path loss, and hence, require advanced processing to ensure a reliable transmission. At the same time, FH and RAN links are closely interwoven from a signal processing perspective by forming a concatenated structure: the inner FH link forwards samples of the outer radio access link.

These facts – a high dependence of FH requirements on future radio access technologies (RATs), the reduced reliability of cost-efficient wireless FH links, and the concatenated structure of radio access and FH – calls for a joint processing of both signals. Such a joint approach is inherently lacking in the current approaches to FH networks, which treat both links as separate network segments, being designed, deployed, and processed with little regard for one another. Current research hence focuses on FH compression to reduce the data rate requirements [CK16a; GCTS13; LC13; SPM+12; PSSS14; PSSS13; ZXCY15] or treats the FH network as a limitation [PWLP15; LPC+14; HLD16; LZ16; ZXCY15]. A first step towards considering the relation between RAN and FH processing currently being investigated, is to reconsider the distribution of the processing between the remote access points and the centralized cloud processors – so-called functional splits [DDM+13; RBD+14; HY+15; lIYH+15]. This thesis investigates methods to more closely couple radio access and FH in order to improve performance and enable a more efficient integration of wireless FH technology into the mobile networks of the future.

1.2 Contribution and Outline

This thesis is organized into eight chapters. Each chapter is introduced by motivating the upcoming investigations and listing the publications of this thesis' author on the particular topics. A summary at the end of each chapter gives a concise overview of the findings. In the following, we outline the contributions of each chapter and highlight the novel approaches developed in this thesis.

- **Chapter 2** lays the groundwork for this thesis. It recaptures the recent progress in the centralization of mobile networks and takes a look ahead to discuss what technologies are required to enable cloud-based 5G networks. Most notably, it analyzes the novel concept of functional splits, which reenvisions the distribution of signal processing between remote radio access points and the central cloud processing centers. While this concept was initially investigated for 4G networks, this thesis extends the concepts to new RATs such as massive multiple-input/multiple output (MIMO) and mmWave communication. The requirements that these novel technologies will induce in the transport network will be analyzed in detail, showing that it is in fact necessary to consider transport limitations already when designing the access technologies for future networks. Finally, mmWave fronthaul is discussed, which is as a promising technology to reduce the costs of fronthaul deployments but suffers from a reduced reliability due to its high pathloss and susceptibility to precipitation. To address this challenge will be the main motivation for the rest of this thesis.

- **Chapter 3** begins a more detailed analysis of the novel concept of joint design of access and FH uplinks, which will be the topic for the rest of the thesis. The concept proposes to more tightly couple reception of FH and RAN signals on the physical layer, which is especially beneficial when using unreliable FH technologies such as mmWave wireless links. For this, Chapter 3 first introduces a detailed system model that will be used throughout the thesis. In contrast to contemporary approaches, this model views FH as a part of the overall RAN channel, hence enabling a joint processing. Basic properties of this joint channel are investigated and it is detailed how this approach can be utilized to improve overall performance.

- **Chapter 4** introduces a novel joint minimum mean-square-error estimation (MMSE) receiver for access and FH signals that incorporates soft information of both signals for improved end-to-end reliability. Several variants of this approach are derived, including variants of different complexity, variants for different functional splits, and an iterative extension of the method. Basic properties and trade-offs of this receiver are presented and it is shown to achieve considerable performance gains over conventional approaches.

- **Chapter 5** takes a closer look at the impact that the quantizer has in the joint RAN/FH system. The quantizer is employed to digitize the radio signal after reception and its design is shown to be of considerable consequence for the overall performance. This impact is investigated and two novel optimization schemes are proposed that result in an improved quantizer yielding further performance increase.

- **Chapter 6** concentrates on investigating the performance of the concepts proposed in the previous chapters within a practical application example. Using a realistic, highly relevant scenario, detailed simulations show the different trade-offs and use cases in which the proposed approaches can be expected to be most beneficial. Clear guidelines are provided, how the proposed concepts are best employed.

- **Chapter 7** investigates the practical aspects of implementing the proposed schemes in real-life, thereby complementing the more theoretic analysis of previous chapters. In this regard, the computational complexity of the required hardware is detailed, and an appropriate, optimized hardware architecture is proposed. Furthermore, aspects of overall network design are discussed, detailing how the approaches can be incorporated in a larger C-RAN architecture.

- **Chapter 8** concludes the technical content of this thesis. It concisely summarizes the main findings and draws conclusions implied by the work presented. It also shows the way ahead by highlighting open points and giving inclinations towards possible future work.

2 Centralization Aspects of Mobile Networks

As this thesis is focused on C-RAN systems, we will first give an introduction to the architecture of such a system, its benefits, requirements, challenges, and current developments. The author of this thesis published several works on this. In [BFW+13], physical technologies are described that can be employed to facilitate the FH and backhaul (BH) networks required for a centralized architecture. A more detailed view on wireless FH together with early ideas about a joint design of RAN and FH networks is presented in [BF13]. The idea of a more flexible split of processing between centralized and decentralized nodes was investigated in [MLD+14]. The virtualization of baseband processing for future 5G networks and the corresponding impact on FH is analyzed in [WRB+14]. The FH and BH requirements resulting from 5G RATs are described in detail in [BRW+15] and [BRW+16].

2.1 Cloud Radio Access Network Architectures

2.1.1 From Distributed to Centralized Radio Access Networks

Historically, mobile networks featured a decentralized signal processing architecture. From the first beginnings until the early deployments of 4G networks, all lower layer signal processing up to the network layer was performed at base stations (BS). A centralized core network was only responsible for general tasks, like keeping a register of subscribers, providing a gateway to other telecommunication networks, or managing mobility across different mobile networks. The actual signal processing however, i.e. transmitting and receiving wireless signals, was performed at each BS individually in baseband units (BBUs). As those BBUs were rather large and heavy, they were deployed at the base of the cell towers, with a digital interface connecting the digital BBUs with analog radio frequency hardware (mixers, amplifiers and antennas) at the top of the tower. Fig. 2.1 shows the structure of a typical decentralized mobile network, which we will refer to as D-RAN in this thesis. The user equipments (UEs) use an access link to communicate with a BS, which are connected to the core network via the so-called backhaul (BH) network. The BH transports user traffic, as well as control, management and monitoring information. Additionally, BSs can also be directly interconnected by BH links to exchange information. Usually, the BH links use packet-switched technology and hence routers and switches are used to route and aggregate traffic. The core network provides the gateways for services like the Internet or fixed line telephone networks. The route from UE to core is referred to as the uplink (UL), while the reverse from core to UE is called the downlink (DL).

The idea of a centralized RAN was motivated by two facts [Chi11]. First, the deployment of BBUs at the base of cell towers requires a large housing or even machine room, which have to be acquired or rented from the original site owner. It was hence considered that it might be more cost-efficient to locate a large number of BBUs in a single location, and simply string out the digital interface to the RF hardware to a longer distance. Second, the development of cooperative techniques like distributed antenna systems (DAS) [CA07] and coordinated multi-point (CoMP) [IDM+11; SGP+13] transmission and reception requires a regular exchange of information between BS, which induces a large amount of BH traffic. A co-location of BBUs simplifies such an exchange significantly. Fig. 2.1 shows a typical centralized RAN system as they are deployed today.

The benefits of a centralized RAN architecture are hence twofold: simplified cooperative processing due to a shorter distance between BBUs, and a smaller size of the hardware to be deployed at the cell site, which are reduced to remote radio heads (RRHs). In addition, the management of the network is simplified, as, for example, broken or outdated baseband hardware can be exchanged in a single location, instead of having to visit separate BS sites. All these benefits are ultimately expected to lead to cost savings for mobile operators.

However, this centralization also brings substantial challenges. First, while the exchange of signals between BBUs is simplified, the antennas are still distributed and their signals need to be transported over the FH network to the BBUs. As will be discussed later, this induces very strong requirements in terms of FH data rate, latency, and synchronization. As these requirements are very challenging to fulfill, dedicated fiber links are commonly used today. These have to be either leased from third parties, or deployed by the operators themselves, requiring extensive civil works and the procuring of rights-of-way. Hence, FH networks are very expensive, thereby mitigating the intended cost savings of centralization [Rea15]. Due to the use of *dedicated* fiber links, FH networks are also very inflexible, as a single fiber channel (i.e., a fiber core or wavelength) has to directly connect the RRH and the BBU, without being able to re-route traffic in case of failure or

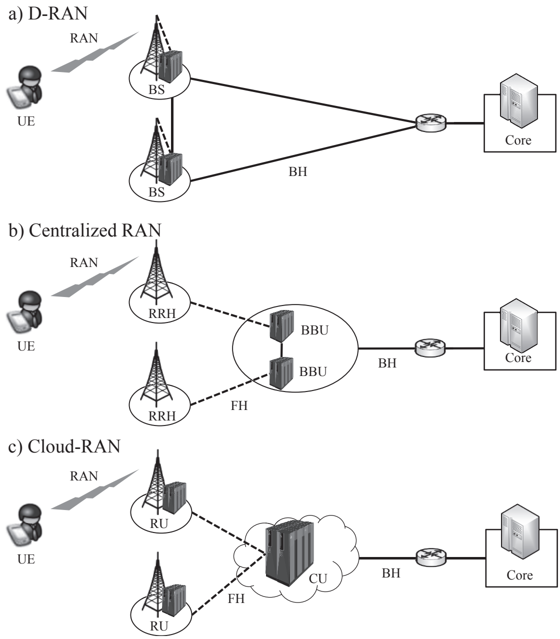

Fig. 2.1.: Comparison of D-RAN (top), centralized RAN (center), and Cloud-RAN (bottom) architecture.

changing network topology, and without being able to share this fiber channel with other services.

2.1.2 The Road to Cloud Radio Access Networks

In order to improve the concept of centralization, several approaches are currently under investigation. First, the introduction of *cloud* technologies, can further improve the cost-efficiency of centralized networks, leading to C-RANs (C-RANs). Cloud technologies were originally introduced in the IT industry and refer to providing a "shared pool of configurable computing resources [. . .] that can be rapidly provisioned" [MG]. The idea behind it is to share processing resources among different users and applications corresponding to their actual demand. It can be expected that in most cases not all users or applications require peak resources. Hence, a lower processing power needs to be deployed overall compared to a scenario when each user/application needs to have resources according to its peak demand provisioned at all times. This approach is considered to be beneficial to mobile networks as well, as peak traffic occurs usually not in all cells at the same time, and hence a lower amount baseband processing resources can be shared among cells, instead of deploying a dedicated BBU for each cell.

This "cloudification" requires a *virtualization* of processing, which refers to a decoupling of physical processing resources from the actual processing algorithms. Instead of using dedicated digital signal processors (DSPs), field programmable gate arrays (FPGAs), or application specific integrated circuits (ASICs) for baseband processing, general purpose processors (GPPs) can be used, on which virtual instances are created to process, e.g., a BS [WRB+14]. The allocation of physical processor cores and processor time to the individual applications is managed by a so-called hypervisor.

This virtualization not only makes it possible to share processing resources, thereby reducing the overall processing power to be deployed. It also reduces the costs of hardware, as commercial-off-the-shelf servers can be utilized instead of specialized baseband hardware, thereby exploiting economics of scale. In addition, the processing algorithms can be modified or updated much easier, as they are implemented merely in software instead of on dedicated hardware. The GPPs could be located in large data centers as used for application processing today, or even shared with other, more "traditional" cloud services like hosting web servers or mass storage, thereby further increasing cost-efficiency.

At the same time, it has been proposed to add a certain degree of general purpose processing power to the edge of mobile networks, i.e. adding small servers to BS sites. These distributed GPPs are referred to as "mobile edge cloud" [CJLF16; DMT+13], "cloudlets" [SBCD09], or "Fog-RAN" [PYZW16; CSK16; CK16b]. While originally intended for user application processing such as augmented reality, this distributed computing power could also be used for baseband processing, further increasing the benefits of shared resources.

The main disadvantage of utilizing GPPs is that they generally offer lower computational power and longer latencies as compared to dedicated hardware, especially for complex baseband operations like fast Fourier transformations (FFT) and turbo decoding [WRB+14; SN06]. At the same time, overhead is introduced via the hypervisor. However, new generations of servers already mitigate these downsides. The approach of so-called "bare metal servers" [GAH+12] remove a part of the hypervisor overhead, by allocating dedicated physical cores to each virtual processes, instead of sharing the processor time among several processes. In addition, servers are being equipped with hardware accelerators like DSPs, that can then be allocated to virtual processes for the more complex operations [Hew].

As the central processing units not necessarily only perform baseband processing, they are referred to as simply central units (CUs), while the units at the cell sites are referred to as remote units (RUs). In fact, the processing can be flexibly assigned to either the CU or the RU, leading to the concept of functional splits, as discussed in the following.

2.1.3 Functional Splits

The term *functional split* refers to a configuration of performing a certain part of baseband processing at the RU, and the remaining part at the CU. First proposed in [DDM+13], the approach of functional splits has received considerable attention in, e.g., [Sma; Rea; RBD+14] and is also a key concept considered in this work. The general concept is that by performing some of the baseband processing at the RU, the requirements on the FH can be reduced. Fig. 2.2 shows a signal processing chain as it is used in today's 4G long term evolution (LTE) systems and can similarly be expected from 5G systems. As this thesis focuses on the UL, the corresponding operations are explained in detail. In the DL, the operations are basically reversed. The following explanations are only intended to give a general overview of the performed processing to enable the reader to follow the subsequent discussion on functional splits. The system model in Sec. 3.1 will provide detailed descriptions of each step as used in this thesis. A detailed description of each step as performed in today's LTE systems can be, e.g., found in [SIM09].

- **Antenna:**

  In UL direction, the signal is first received at the antenna. The antenna provides conversion from the radio wave to an electric current in the receiver circuit. It can also add directivity and an antenna gain, thereby increasing the received energy. In addition, multiple antennas can be combined to either provide additional antenna gain, directivity or spatial diversity. While currently up to 8 antennas are used at RUs, this number is expected to rise for 5G to hundreds of antennas to utilize massive MIMO technology [LETM14].

  In general, the antenna gain is coupled to the directivity of the antenna: by focusing the transmitted energy into a certain direction, the received power is increased. This process is also referred to as beamforming. The gain and directivity depend on the shape of the antenna, and on the ratio of the antennas aperture and the utilized wavelength. In that regard, higher

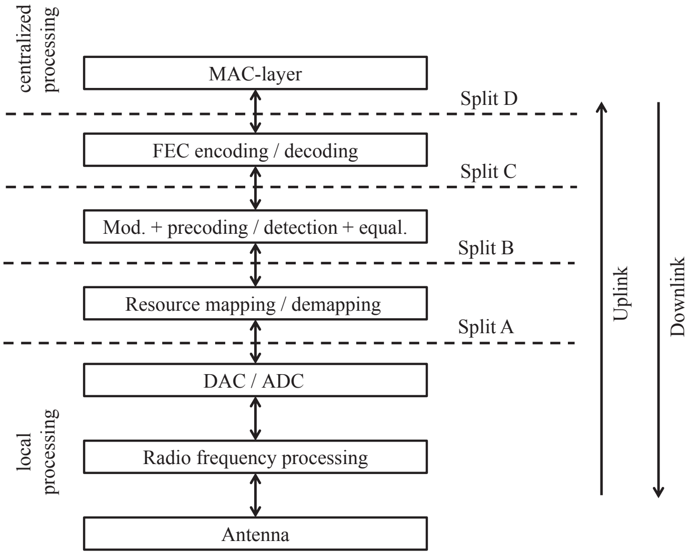

Fig. 2.2.: Signal processing chain and functional split options.

frequencies are advantageous, as either antennas can be built smaller or can achieve a higher gain at the same aperture size.

- **Radio Frequency Processing:**
  In general, the high-frequency radio signal is first filtered to select the appropriate channel, amplified and then down-converted to a baseband signal. However, many different receiver architectures with different stages of filtering and mixing exist, which shall be not discussed in detail here. An overview of analog receiver structures can be found, e.g., in [MUM07]. In addition to the traditional RF reception, the analog signals of multiple antennas can be combined for beamforming at this stage, by, e.g., using a combiner network with phase shifters [AMGH14], or using a lens antenna array [ZZ16].

- **ADC/DAC:**
  Next, the signal is analog-to-digital (A/D) converted by sampling and quantization. The sampling rate is usually selected according to the channel bandwidth, while a certain degree of oversampling is usually performed to improve signal quality and foster easier FFT implementation. The quan-

tization resolution is selected to keep quantization noise to a tolerable level.

- **Resource Demapping:**
  The received samples of a digital transmission contain multiple different signals. First, multiple users are usually multiplexed in time and/or frequency. In the same fashion, additional signals for synchronization, channel estimation, and control are multiplexed. These signals are separated in this step. In LTE, an FFT is performed first, as LTE utilizes single-carrier frequency division multiple access (SC-FDMA).

- **Equalization and Detection:**
  As the signals in mobile networks are impaired by multi-path propagation and the Doppler effect, the channel needs to be equalized next in order to compensate frequency- and time-selective attenuation and phase shifts. With the help of aforementioned reference symbols, this can be performed in either time or frequency domain. Due to the much simpler implementation, frequency domain equalization (FDE) is used in today's LTE systems. In addition, the signals of multiple antennas can be combined at this stage via digital MIMO processing.

  During the following demodulation, the data signals are detected according to the used modulation scheme to recover the transmitted bits. While this can be performed as hard decisions based on thresholds, modern receivers utilize so-called soft detectors, which calculate reliability information on the individual bits based on the distance of the received symbols to the transmit symbols.

- **Decoding:**
  To correct erroneously detected bits, the original user data is protected by a forward error correcting code (FEC), e.g. a turbo code [Skl97] is utilized in LTE. As the FEC can only correct errors up to a certain degree, a cyclic redundancy check (CRC) code is added first, which basically provides a checksum to judge the integrity of the data. In the receiver, the FEC is decoded, and based on the CRC the success of the decoding operation is decided. This decision serves as input for the hybrid automatic repeat request (HARQ) process, i.e., the successful reception is either acknowledged (ACK'ed) to the UE, or a repetition of the transmission is requested with a negative acknowledgment (NACK).

- **MAC Layer:**
  After the HARQ processing, the user data is handed over to the medium

access control (MAC) layer, where the data is passed to radio link control (RLC), the packet data conversion protocol (PDCP) and radio resource management (RRM) layers. Conventionally, the HARQ process is already counted as part of the MAC layer. However, as it is an inherent part of the decoding process, it will be described together with physical layer (PHY) processing within this thesis.

This signal processing chain can in principle be split between any two of the above described steps, or even within one of the steps, as they usually are comprised of multiple operations. However, not every option is feasible in practice or brings substantial benefits. Hence, four representative splits are discussed in the following to illustrate the basic trade-offs of splitting the processing between RU and CU. These split options are referred to as A, B, C, and D, although other notations have been used by different groups of authors according to the number of splits discussed, e.g. in [Sma; WRB+14]. The splits will be introduced shortly in the following, while their specific requirements will be discussed in detail in Sec. 2.2.

- **Split A:**
  The first option is to split processing immediately after A/D conversion. The data forwarded to the CU in this case is raw in-phase/quadrature-phase (I/Q) samples. The main advantage of this split is that all digital processing is located in the CU, and hence, the RU can be built in a small and lightweight form factor, thereby simplifying deployment. In addition, any type of cooperative processing such as CoMP can be performed jointly in the CU, without having to exchange large amounts of data between cell sites. The main disadvantages of this option are, as will be discussed in the next section, strict requirements on latency and synchronization, as well as the high and utilization-independent demand in FH capacity due to the redundancy included in the raw I/Q samples.
- **Split B:**
  For split B, the different signals – data, reference symbols, synchronization signals, guard carriers, and control information – are demapped at the RU. The data forwarded are still I/Q samples. However, in case of multi-carrier waveforms, the samples would now be in the frequency domain instead of in the time domain. More importantly, redundant signals can be discarded, which includes guard carriers, cyclic prefixes, and all data resources that are not currently utilized. As data is only transmitted when users want to communicate, resources are left unused when the demand is low. In split

A, these unused resources are forwarded as well, while after the resource demapping this is not the case. As will be discussed in the next section, this also leads to a *varying* FH traffic for split B, as compared to the *static* traffic of split A.

In addition, synchronization and channel measurements could already be performed at the RU, in which case the corresponding signals also do not have to be forwarded over the FH, only the processing result (e.g. channel state information).

- **Split C:**

  For split C, channel estimation and symbol detection are also performed at the RRH. The demodulator output is comprised of log-likelihood ratios (LLRs) corresponding to individual bits. Hence, the data rate is coupled to the utilized modulation scheme, i.e. for 4-QAM (quadrature amplitude modulation), two LLRs per symbol need to be forwarded, for 16-QAM four LLRs, and so on. The modulation scheme in turn is selected according to the UEs channel quality. This means, that the FH traffic is even tighter coupled to the users' data rate: not only is no traffic forwarded when there is no user demand, but when a user faces bad channel conditions and receives a lower throughput, this will be also reflected in the FH traffic.

  Since channel equalization is now performed at the RU, no reference symbols need to be forwarded to the CU. However, this also removes the option of employing joint detection methods like CoMP's joint reception [SGP+13] centrally, requiring the necessary information to be exchanged between cell sites, which nullifies one of the main advantages of a centralized RAN architecture. Hence, centralization benefits can only come from cooperative decoding, or higher layer functionalities like CoMP's joint scheduling [SGP+13] or resource allocation.

  In addition, the signals of different antennas are combined during MIMO processing in this step. Hence, the data rate depends on the number of spatial layers from now on, instead of on the number of ADC chains.

- **Split D:**

  Split D in addition centralizes the FEC decoding, thereby allocating all PHY layer processing to the RU. This removes the redundancy added to the user data to compensate for transmission errors. As the code rate, i.e. the amount of redundancy, is again selected based on the UEs' channel quality, this further increases the coupling of FH traffic and the throughput experienced by users. In that regard, the traffic of split D is in fact more similar to conventional BH traffic. The data forwarded is now in the form of hard bits and is dominated by the actual user traffic, with some additional headers and control information.

As these splits are fundamentally different, each comes with a distinct set of requirements that need to be met by the FH. The introduction of 5G RATs will also have a considerable impact on these requirements, which will be discussed in the next section in detail.

2.2 Requirements of Future Transport Networks

2.2.1 5G Radio Access Technologies

The requirements of the transport network are dictated by the requirements of the RAT, which in turn is designed based on the considered applications. While LTE as a baseline technology is well understood regarding its FH requirements, new technologies will be introduced for 5G. Hence, in order to understand the FH requirements of the future, the most popular technologies under consideration for 5G are shortly summarized first. These include:

- Carrier aggregation (CA) [YZWY10] uses the combination of multiple LTE channels to achieve higher throughput.
- Higher bandwidth channels are considered to increase the data rate, e.g. in [3GP16a]. This could require the allocation of additional bands in the sub-6 GHz range.
- The allocation of mmWave bands [RSM+13] could increase the available bandwidth even more drastically, enabling channels of multiple GHz. However, the pathloss increases at higher frequencies, hence requiring either smaller cells or higher gain antennas.
- Massive MIMO technologies [LETM14] introduce antenna arrays with a large number of antenna elements. This can be used both for a higher antenna directivity as well as spatial separation to either overcome pathloss or to increase data rates.
- Higher order modulation schemes, such as 256-QAM are already introduced in the latest releases of LTE [KBNI13] and could be further increased for 5G.

- The introduction of new waveforms is considered [WJK+14] in order to use the spectrum more efficiently by reducing the overhead in cyclic prefixes and guard carriers. They can also facilitate a more flexible definition of resource grid configuration to better adapt to new applications.
- The introduction of the Tactile Internet [Fet14] requires a greatly reduced end-to-end latency.
- New applications such as factory automation [FWB+14] or vehicular communication [SPS15] require a very high reliability.
- Machine-type communication (MTC) [RPL+15] will result in an dramatically increased number of connections, albeit with low average traffic per device.

Having discussed 5G RAN technologies, the requirements for future transport networks can be derived. For this, the FH technologies already standardized for 4G can be used as a baseline. The most commonly used standard for fronthauling utilized today is the Common Public Radio Interface (CPRI) [Com]. It implements split A, i.e. all digital processing is performed in the CU. The standard also defines requirements to be fulfilled by the FH links to ensure performance on the RAN link. These requirements most prominently include latency, data rate, delay estimation accuracy, jitter, and reliability. In order to illustrate these requirements, three exemplary 5G RATs are being analyzed in this sections. The main parameters of these systems are summarized in Table 2.1.

The three systems are mainly differentiated by their carrier frequency. The first system uses a carrier frequency of 2 GHz, and hence represents an evolutionary step from 4G LTE. To reflect a higher capacity demand, five-fold carrier aggregation is assumed, as well as the recently introduced 256-QAM. In addition, a large antenna array with 96 antennas is assumed for enhanced MIMO, with up to 8 spatial layers. The transport overhead of 33 % corresponds to one control bit per 15 payload bits together with an 8b/10b line coding [WF83] as utilized in today's CPRI. Similarly, the quantizer resolution is chosen as 15 bits for split A. As it was observed in [DDM+13] that the resolution can be reduced in the frequency domain, 12 bits[1] are chosen for split B. For split C, LLRs are quantized, so three quantization bits are used, which is sufficient for good performance [DFA+10]. As for split D user bits are forwarded that do not need to be quantized further, the number of bits is given as one. The remaining parameters (maximum code

[1] 7-9 bits are given as sufficient in [DDM+13], however, only-64 QAM was considered therein. To account for the lower inter-symbol distance when using 256-QAM, 12 bits are assumed here.

rate, subframe duration, FFT size, number of subcarriers, reference overhead and number of symbols) reflect a typical 20 MHz LTE configuration. The maximum utilization is 100 %, i.e. full load.

The second system employs a 30 GHz carrier, thereby representing one of the lower mmWave bands currently under consideration. According to the higher available bandwidth, a 500 MHz channel is assumed. Due to the smaller wavelength, a larger number of antennas can be assumed. However, this large number of antennas is mainly required to overcome the increased pathloss. The increased directivity also limits multi-path propagation and hence the number of potential spatial layers, which was hence chosen as a maximum of four. The FFT size was kept identical to the 2 GHz system for easier comparison, and the frame duration and number of symbols per frame were adapted according to the higher bandwidth. The chosen structure is also based on a multi-carrier system, although single-carrier systems might be employed for mmWave carriers. Note that the exact frame structure and number of subcarriers have no impact on the final data rates in this section as the number of samples per time interval remains constant. As the ADCs operate at a higher sampling rate, it can be expected that the quantizer resolution has to be reduced to limit power consumption, so 9 bits are used for split A and B. Accordingly, the modulation order is assumed to be limited to 64-QAM.

The third system uses a 70 GHz carrier, i.e. a higher mmWave bands. The differences to the 30 GHz system include a higher channel bandwidth of 2 GHz, a further increased number of antennas, and a slightly adapted frame structure.

Since the 5G standardization is not finished at the point of writing of this thesis, the parameters will most likely change in the future. However, they already give a good impression of the principle effects that the discussed technologies will have. Some early standardization from 3GPP is available in [3GP16a]. This technical report considers also carriers below 6 GHz, and around 30 GHz and 70 GHz, a bandwidth of up to 1 GHz, and up to 256 antennas. The values chosen in Table 2.1 are hence well in line with what can be expected from future standardization.

Tab. 2.1.: Exemplary parametrization and requirements of 5G RATs.

Parameter	Symbol	Unit	Sub-6 GHz	Low mmWave	High mmWave
Carrier frequency	f_{C}	GHz	2	30	70
Channel bandwidth	BW	MHz	100	500	2000
Sampling rate	f_{S}	MHz	150	750	3000
Antennas	N_{A}	-	96	128	512
ADC/DAC chains	N_{P}	-	8	4	4
Max. spatial layers	N_{L}	-	8	4	4
Transport overhead	ζ	-	1.33	1.33	1.33
Quant. res. Split A	$N_{\mathrm{Q,A}}$	Bits	15	9	9
Quant. res. Split B	$N_{\mathrm{Q,B}}$	Bits	12	9	9
Quant. res. Split C	$N_{\mathrm{Q,C}}$	Bits	3	3	3
Quant. res. Split D	$N_{\mathrm{Q,D}}$	Bits	1	1	1
Max. modulation order	m_{RAN}	-	256	64	64
Max. code rate	$R_{\mathrm{C}}^{\mathrm{RAN}}$	-	0.85	0.85	0.85
Subframe duration	T_{S}	ms	1	0.1	0.1
FFT size	N_{FFT}	-	2048	2048	2048
Active subcarriers	N_{SC}	-	1200	1200	1200
Reference overhead	η	%	10	10	10
Symbols per frame	N_{Sy}	-	70	35	120
Max. utilization	μ	%	100	100	100

2.2.2 Data Rate

The data rate depends heavily on the split employed and in general is lower for lower degrees of centralization. In the following, formulas for each split are given, with the symbols explained in Table 2.1. Similar formulas are derived in [DDM+13; BRW+15; WRB+14]. The formulas will be first discussed in general, before numerical examples will be given towards the end of this subsection.

Split A

The data rate for split A can be given as

$$D_{\mathrm{A}} = \underbrace{N_{\mathrm{P}}}_{\text{Antenna ports}} \cdot \underbrace{f_{\mathrm{S}}}_{\text{Sampling freq.}} \cdot \underbrace{N_{\mathrm{Q,A}}}_{\text{Quant. res.}} \cdot \underbrace{2}_{\text{I/Q}} \cdot \underbrace{\zeta}_{\text{FH overhead}} . \tag{2.1}$$

As in split A the received signal is only sampled and quantized, the data rate is static and depends on the sampling frequency, which is chosen according to the channels bandwidth. In Table 2.1, an oversampling factor of 1.5 was chosen, as used in LTE[2]. The introduction of carrier aggregation as well as higher bandwidths available at mmWave frequencies linearly scales the required FH data rates, which hence have to be expected to be much higher for 5G networks.

Furthermore, the data rate of split A depends on the number of ADC chains. This is an important point, as CPRI specifies one data stream per *antenna*. With the introduction of massive MIMO with potentially hundreds of antenna elements, the data rate would increase dramatically. It is hence necessary to perform beamforming already at the RU, and only forward a reduced number of streams according to the maximum number of spatial layers. This is especially important to consider for mmWave carriers, as they will have to rely on large antenna arrays to overcome the increased pathloss, yet at the same time will face lower spatial diversity due to less prominent multi-path effects in the directive links.

In addition, the data rate of split A depends on the resolution of the A/D converter, with the factor 2 accounting for the I and Q phase being quantized separately. While current systems use relatively high resolutions of around 15 bits to account for the high dynamics of the signal, lower resolutions will have to be applied for mmWave frequencies, due to the increased power consumption of high-rate ADCs [LS08].

[2]Note that the exact sampling frequency used in 20 MHz LTE systems is 30.72 MHz. The rounded value was used for simplicity.

The FH overhead factor accounts both for control information that is transmitted along with the I/Q data, and additional coding. Fiber links usually employ either 8b/10b or 64b/66b line coding. As wireless links are more prone to error as compared to fiber, additional overhead in terms of a FEC potentially has to be added for such a case.

As will be shown below, split A exhibits a very high data rate. The main reason for this is the oversampling and forwarding of raw, unprocessed samples. This high FH data rate is the main reason for the high deployment costs of current, CPRI-based FH networks.

Split B

The data rate for split B is given as

$$D_{\mathrm{B}} = \underbrace{N_{\mathrm{P}}}_{\substack{\text{Antenna}\\ \text{ports}}} \cdot \underbrace{N_{\mathrm{SC}}}_{\text{Subcarriers}} \cdot \underbrace{N_{\mathrm{Sy}}}_{\text{Symbols}} \cdot \underbrace{T_{\mathrm{S}}^{-1}}_{\substack{\text{Frame}\\ \text{duration}}} \cdot \underbrace{\mu}_{\text{Utilization}} \cdot \underbrace{N_{\mathrm{Q,B}}}_{\substack{\text{Quant.}\\ \text{res.}}} \cdot \underbrace{2}_{\text{I/Q}} \cdot \underbrace{\zeta}_{\text{FH overhead}} . \tag{2.2}$$

Instead of depending on the sampling rate, the data rate now depends on the product of the number of subcarriers N_{SC}, the number of symbols per subframe N_{Sy} and the inverse of the subframe duration T_{S}. Note that this product reflects the number of samples per given timeframe. However, this time the data is not oversampled. After FFT conversion to the frequency domain, guard carriers can be removed, which corresponds to a downsampling with a lower rate in time domain. To give a practical example, a conventional 1.4 MHz LTE system uses an oversampling of 1.92 MHz. From these samples, first the cyclic prefix is removed, with the remaining samples used for a 128-point FFT. Of the 128 subcarriers now available, 56 are guard carriers which can be discarded after the FFT. The remaining 72 subcarriers now represent a bandwidth of only 1.08 MHz. This is illustrated in Fig 2.3, where the different signals in the resource grid are shown. As can be seen, a considerable part of the samples can be discarded for split B, thus reducing the FH data rate dramatically.

In addition, it was observed, that a lower number of bits $N_{\mathrm{Q,B}}$ can be used for the frequency domain representation [DDM+13], thereby further reducing the data rate.

The most important factor however, is that only resources occupied for user data transmission according to the current cell load have to be forwarded. In Eq. (2.2), this is represented by the utilization factor μ. The importance of this load dependence will be analyzed in more detail later in this section.

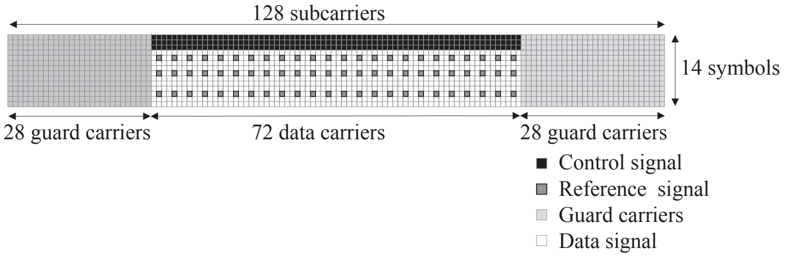

Fig. 2.3.: Simplified LTE resource grid illustrating ratio of data symbols (white squares) to other symbols (gray/black) for 1.4 MHz configuration ($f_S = 1.92$ MHz, $N_{FFT} = 128$).

Split C

The data rate for split C can be calculated as

$$D_C = \underbrace{N_L}_{\text{Layers}} \cdot \underbrace{N_{SC}}_{\text{Subcarriers}} \cdot \underbrace{N_{Sy}}_{\text{Symbols}} \cdot \underbrace{T_S^{-1}}_{\substack{\text{Frame}\\\text{duration}}} \cdot \underbrace{\mu}_{\text{Utilization}} \cdot \underbrace{\log_2(m_{RAN})}_{\substack{\text{Modulation}\\\text{scheme}}} \cdot \underbrace{(1-\eta)}_{\substack{\text{Reference}\\\text{overhead}}} \cdot \underbrace{N_{Q,C}}_{\substack{\text{Quant.}\\\text{res}}} \cdot \underbrace{\zeta}_{\text{FH overhead}}. \tag{2.3}$$

As the equalization and detection step includes MIMO processing, the dependence on the number of antenna ports is replaced by the number of spatial layers. While a cell might in theory support a high number of different antenna streams, the channel conditions might not yield enough spatial separation for independent streams. In such cases, the different antenna ports can still be combined for receive diversity at this stage.

The data forwarded after this step will consist of LLRs for each of the transmitted bits, and is hence coupled to the modulation scheme m_{RAN} utilized for each symbol. Hence, a corresponding factor is introduced in Eq. (2.3), as well as an appropriate quantizer resolution of three bits for each soft value. In addition, reference symbols (dark gray squares in Fig. 2.3) for channel estimation are no longer required to be forwarded, hence the amount of forwarded resources is again reduced by a corresponding factor $1 - \eta$.

Split D

The data rate of split D can be calculated as

$$D_D = \underbrace{N_L}_{\text{Layers}} \cdot \underbrace{N_{SC}}_{\text{Subcarriers}} \cdot \underbrace{N_{Sy}}_{\text{Symbols}} \cdot \underbrace{T_S^{-1}}_{\substack{\text{Frame}\\\text{duration}}} \cdot \underbrace{\mu}_{\text{Utilization}} \cdot \underbrace{\log_2(m_{RAN})}_{\substack{\text{Modulation}\\\text{scheme}}} \cdot \underbrace{R_C^{RAN}}_{\substack{\text{Code}\\\text{rate}}} \cdot \underbrace{(1-\eta)}_{\substack{\text{Reference}\\\text{overhead}}} \cdot \underbrace{N_{Q,D}}_{\substack{\text{Quant.}\\\text{res}}} \cdot \underbrace{\zeta}_{\text{FH overhead}}. \tag{2.4}$$

As the output of the decoder are information bits, the quantizer resolution here is always one bit. In addition, the redundant bits added for error protection are removed, leaving only the original payload bits. Hence, the FH data rate is reduced according to the code rate $R_{\mathrm{C}}^{\mathrm{RAN}}$. The quantizer resolution is now equal to one bit since hard bits are forwarded. As mentioned before, this split corresponds closely to a classical BH split, and the data rate is very close to the actual throughput perceived by users.

Peak Data Rates

Having derived equations for the data rate of the individual splits, the implications for real networks will be discussed next. First, the peak data rates for all three splits and the three different RATs from Table 2.1 are illustrated in Fig. 2.4.

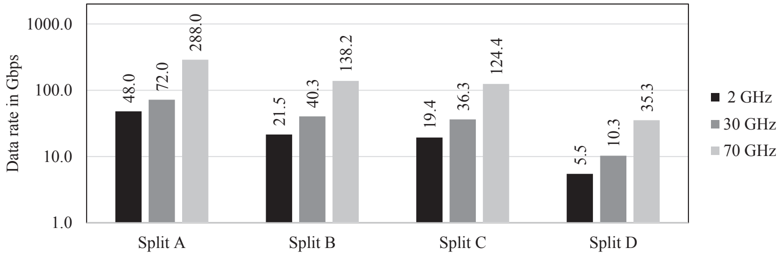

Fig. 2.4.: FH throughput requirements for four different functional splits and three RATs with $f_{\mathrm{C}} = \{2, 30, 70\}$ GHz.

Two important observations can be made: first, the required data rates are much higher than in today's networks, where it ranges from 10 Gbps for split A (CPRI line rate 7, 8 antennas, 20 MHz) to 200 Mbps for Split D (BH rate for 2x2 MIMO, 20 MHz). This increase is expected as the systems are designed to support a higher user traffic. This user traffic is represented by split D, which is close to traditional BH. As can be seen, the user data rate would range from 5.5 to 35.3 Gbps, which is approximately in line with the current requirements considered for 5G in, e.g., [NGMb]. Second, the data rates reduce dramatically from split A to D due to the processing performed at the RU. The difference is up to factor 8, which indicates the high price that is to be paid in form of FH capacity for a high degree of centralization.

Variable and Aggregated Data Rates

The values given in Fig 2.4 represent only the peak data rates of the splits. However, it was noted earlier that a main advantage of splits B to D is that they are more closely coupled to the actual user traffic, i.e. their traffic will be lower in times of low demand or when the UEs face unfavorable channel conditions. This fact can be exploited via statistical multiplexing, which occurs, when varying traffic of several cells is aggregated at certain points in the transport network. In order to ensure that all traffic can be transported, the capacity of the aggregation links could be dimensioned for peak traffic. However, if the individual cells' traffic is varying, then dimensioning for peak traffic is highly ineffective, as it is very unlikely that all cells will exhibit peak traffic at the same time. Hence, it is common practice, to dimension the aggregation capacity only for a certain percentile of the traffic, e.g. to be able to transport the traffic with a probability of 95 % [dFV07]. In the remaining 5 % of the cases, this would lead to outage, e.g., increased latency, reduced throughput, or packet loss. However, this is acceptable to operators as it can reduce the required capacity dramatically, which is illustrated in Fig. 2.5. It shows the aggregated probability density functions (PDFs) of 1, 2, 4 and 8 cells when assuming a uniform load distribution between 0 and 100 % in all cells, and for the case of split B (in which case the data rates only vary with the load). The data rates can now be perceived as random variable which are added at the aggregation node. If we assume that the traffic in the cells is independent, this leads to a convolution of the PDFs, which according to the central limit theorem converges to a Gaussian distribution [Ric06]. The 0 % and 5 % outage rates, (corresponding to the 100th and 95th percentile of the distribution) are now given as 172 Gbps and 113 Gbps, respectively. Thus, the required capacity can be reduced by approximately 34 % by accepting a certain outage. Of course, lower outage probabilities can be chosen at the cost of less multiplexing gain.

The statistical multiplexing will be a key aspect of future transport systems and is one main motivation for utilizing new functional splits. Due to the static data rate required for CPRI, statistical multiplexing is currently not possible and the FH network has to be always dimensioned for peak capacity. For splits B to D, however, the transport traffic is increasingly coupled to the actual user traffic and hence, more variable, opening the possibility for multiplexing. To illustrate the effect of these different dependencies, more diverse scenarios than the previous simplified example need to be evaluated. For this, measurements from a real-life network will be analyzed next. The network consisted of 33 LTE

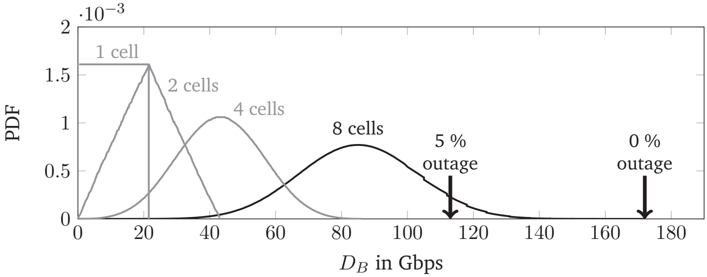

Fig. 2.5.: Aggregated data rate PDFs of different number of cells for split B, 2 GHz RAT and uniform load distribution, and corresponding outage rates.

cells in a major European city, and the measurements were taken over a period of 15 days with an interval of 15 minutes. Fig 2.6 shows the corresponding variation in MCS and load that were observed. The top of Fig 2.4 shows the distribution of the 28 different modulation and coding schemes (MCSs) in each 15-minute interval, and the bottom shows both the average utilization and the maximum utilization among all 33 cells. The loads have been scaled to reflect the higher traffic demand of 5G systems; more details can be found in [5GX16]. Fig 2.7 in addition shows the complementary cumulative distribution function (CCDF) of the loads, both for all time intervals, as well as for the busy hour. The busy hour was selected as the hour with the highest average load, which was found to last from 12:15 h to 13:15 h. What is most interesting here, is that most cells exhibit only low loads during most times, and only in a few time instance cells are highly loaded. Even in the busy hour, the cells are loaded more than 50 % with a probability of only 7.5 %.

The variability of the traffic of the different splits depends on the load (split B to C), on the modulation scheme (split C to D) and the code rate (split D). This is illustrated in Fig. 2.8. It shows the CCDF of the FH data rates of the 2 GHz RAT for split A to D, once using full load in all cells, and once using the variable loads according to the busy hour. As can be seen from the figure, the rate for split A is always constant as previously discussed. Split B depends only on the load and is hence constant for full load. Split C depends on the modulation scheme, and hence it exhibits four values in the case of full load according to the modulation schemes 4-QAM, 16-QAM, 64-QAM, and 256-QAM. Split D finally depends on the combination of modulation and coding scheme, and hence exhibits 29 steps according to the 29 MCS utilized in this example. For the variable load, splits B to D in addition vary according to the load distribution.

In order to extrapolate these results to arbitrarily large networks, the distribution derived from these measurements are next assumed to reflect the probability distribution of the traffic in *any given cell*. In other words, the variations among the measured 33 cells are assumed to be exemplary for each cell individually. By convolving the resulting PDF, the aggregated data rates of arbitrary number of cells can be obtained, similar as was shown in Fig. 2.5 for the case of a uniform distribution. The resulting gain from statistical multiplexing is illustrated in Fig.2.9. It shows the capacity required to aggregate a certain number of cells for the four different splits and the 2 GHz RAT, for three different dimensioning approaches: when dimensioning for 0 % outage, i.e. peak capacity, when taking 5 % outage capacity of a *single cell* and scaling it by the number of cells (this is neglecting the statistical multiplexing), and when taking the 5 % outage capacity of the *aggregated cells*. Since statistical multiplexing cannot be relied on for a low number of aggregated cells, the maximum of the described capacities and the single cell peak rate is taken, which results in the constant regions visible for low numbers of cells. Such an approach was chosen as it is recommended by NGMN in [NGMa].

From the figure, the benefit of accepting outage can be observed as the difference between the dotted and dashed lines, and the additional benefit of

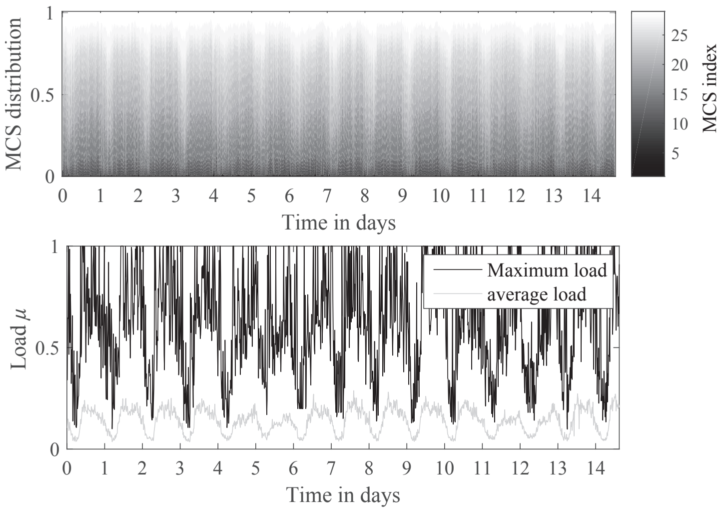

Fig. 2.6.: Measured MCS distribution (top) and scaled average and maximum load (bottom) of 33 real-life LTE cells over 14 days.

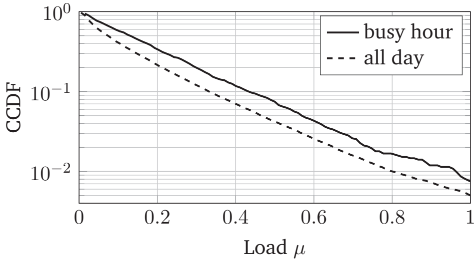

Fig. 2.7.: Busy hour and overall load CCDF based on measurements from 33 real-life LTE cells over 14 days.

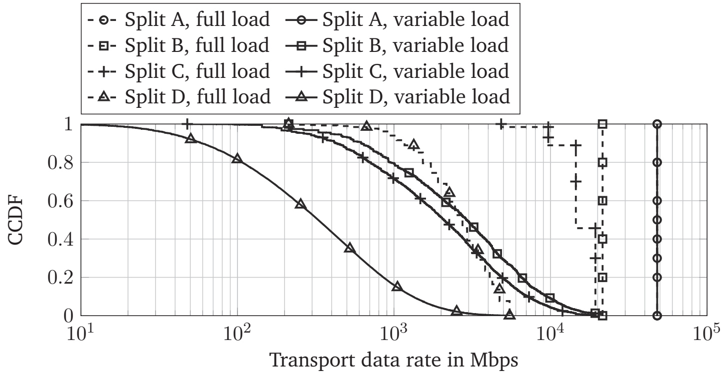

Fig. 2.8.: FH data rate CCDF for 2 GHz RAT and all four functional splits, both for full load (dashed lines) and variable busy hour load according to Fig. 2.7.

multiplexing several cells as the difference between the dashed and solid lines. As the data rate of split A is static, all three curves are the same for split A. As can be seen, the FH capacity is reduced up to factor 10 when both considering outage and statistical multiplexing, showing the benefit of the lower splits B to D. Similar results on statistical multiplexing were observed in [Che16], where statistical multiplexing gains of up to factor 6 were reported based on a different model.

In summary, the above analysis shows that the transport data rates will strongly increase in 5G systems. Especially split A in combination with a mmWave RAT requires tremendous amounts of capacity. In fact, the required data rate of 288 Gbps per 70 GHz cell (cf. Fig. 2.4) could render a fully centralized mmWave RAT infeasible. Several works have hence aimed to decrease the required data rates by introducing I/Q compression. The redundancies contained in the raw

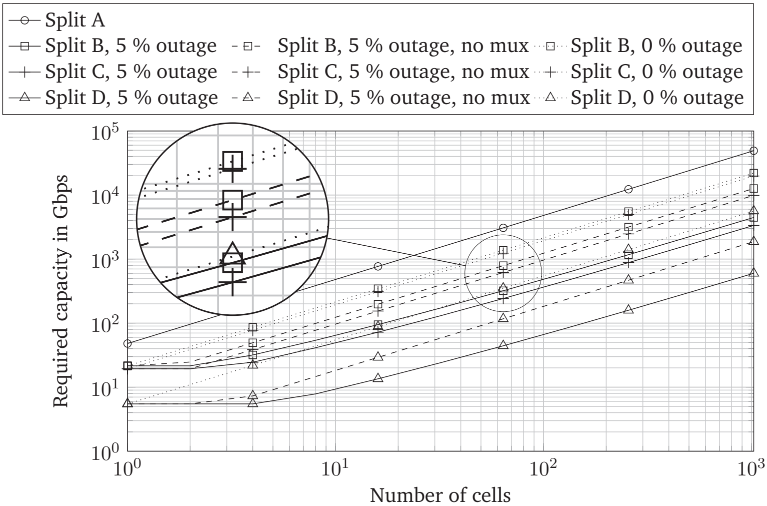

Fig. 2.9.: Required aggregated FH capacity over number of aggregated cells for 2 GHz RAT, all functional splits and for 0 % outage (dotted lines), 5 % outage without considering multiplexing (dashed lines) and 5 % outage when considering multiplexing (solid lines).

I/Q samples can be removed without modifying the functional splits. Works such as [GCTS13; SPM+12; LC13] have proposed compression schemes for CPRI to achieve compression ratios between 1:2 and 1:7. Other works (see [PSSS14] and references therein) have proposed to exploit correlation among the signal of multiple cells in joint compression/decompression schemes. However, as was shown here, statistical multiplexing can decrease the required data rate by the same order of magnitude without relying on elaborate joint schemes. Hence, functional splits are a strong candidate technology to by employed for 5G transport.

2.2.3 Latency

Latency (sometimes also referred to as delay) refers to the time that is required to process a signal for and transmit it over the FH. Additional time is required to process and transmit the signal via the RAN link, and to forward and process it in an application server. However, the latency discussed in the following is specifically the additional time spent in C-RAN systems as compared to D-RANs.

CPRI Limitation

The CPRI standard defines a maximum delay of 5 μs, however this excludes the propagation time on the FH medium, which can be considerably higher (e.g., 50 μs/10 km fiber propagation). For practical purposes, the maximum delay in LTE is limited by its HARQ scheme [CPC+13]. In LTE, UEs can retransmit packets if the BS informs them that a packet was not decoded successfully (NACK). For this, they buffer transmitted packets, but only up to 8 ms. Hence, the time to transmit a packet, process it at the BS, send an ACK or NACK back to the UE and process it there must take less than 8 ms. Subtracting the time typically required for processing and air interface transmission, about 200 μs are left for FH transmissions [SS].

However, this practical limitation is induced by the standard, not by physical limitation. Hence, several methods have been proposed to remove the dependency of the FH latency on LTE's HARQ scheme. In [RP14], an opportunistic HARQ method was proposed, for which the successful transmission of a packet was estimated based on the overall signal quality, without waiting for the packet to be decoded. The event of a packet being wrongly ACKed is handled by higher layer ARQ schemes. While this can increase the FH's delay budget, it leads to a small decrease of overall throughput.

Another approach, called suspended HARQ, is proposed in [DDM+13]. For this, the fact is exploited that UEs clear their buffer even after 8 ms only if the BS schedules the UE to transmit new data. Hence, a forced ACK can be sent without waiting for the decoding outcome, and simply not scheduling new UL transmission before decoding is finished. As no new UL transmissions can be scheduled before the ACK is calculated, the throughput of the UE is reduced. However, since usually multiple UEs share a single BS's resources, another UE can utilize the unused time, thereby interleaving the UEs' HARQ processes.

The most important observation of the latency constraint however, is the fact that a limitation is induced on the *FH network* by the standardization of the *RAN*. In view of ongoing standardization for 5G networks, such an issue can be avoided by simply standardizing UEs to support a longer buffering period. This fact illustrates the interdependence of RAT and FH technologies, and shows that a *joint design* of RAN and FH is required, i.e. considering the implications that one network segment has on the other already when designing 5G technology. This concept will be further extended in Sec. 3.3. Further limitations result either from the users' application, or from the channel coherence time.

Limitation by User Application

The main applications considered for 4G networks were web browsing, video streaming, and voice. As given by [Fet14], the human reaction time – and hence tolerable latency – is on the order of 1 s (browsing) and 100 ms (voice), and can be even higher for video streaming due to the implemented buffering. Yet, for 5G applications such as gaming and the Tactile Internet, a tolerable latency on the order of only 10 ms to 1 ms has to be considered [Fet14]. However, the given latencies are end-to-end latencies, of which only a part can be spent on the FH, with the rest being required for RAN processing and transmission, as well as for the application itself. Hence, the FH latency can only be a fraction of these latencies.

Limitation by Channel Coherence Time

More interesting in the scope of this thesis is the limitation by the channel coherence time. It is well known, that mobile channels exhibit time fading, i.e. they change over time due to the Doppler effect induced by users' movements. In addition, the users' movement changes the overall channel attenuation by a change in distance to the BS and by objects blocking the transmission path. These two effects are referred to as small-scale and large-scale fading. Small scale fading can be compensated by channel estimation in combination with appropriate precoding and equalization, while large scale fading is compensated by adaptive coding and modulation (ACM), as well as gain control. All of these methods require channel state (CSI) or channel quality information (CQI) to be available at the receiver. However, the performance will decrease if this information is outdated. If the information is required in the CU, then the FH latency leads to outdated information. To give an example, the channel state is measured in an UL transmission and then used in the CU for DL precoding. By the time the precoded signal is transmitted from the RU, the FH had to be used twice (once for transmitting the CSI from RU to CU, once for forwarding the precoded signal from CU to RU), and hence the channel can have changed dramatically.

Overall Fronthaul Latency

A simple formula for the channel coherence time is given in [Rap02] as

$$T_{\mathrm{c}} = \sqrt{\frac{9}{16\pi}} \cdot \frac{v_{\mathrm{light}}}{v_{\mathrm{UE}} \cdot f_{\mathrm{C}}}, \tag{2.5}$$

with v_{light} being the speed of light, f_{C} being the carrier frequency, and v_{UE} being the UE's speed.

The latency on the FH is comprised of a processing time, including processing at the RU, the CU as well as intermediate nodes such as switches, and a propagation time. When comparing this to channel coherence times, the additional RAN processing, time, i.e. time for RAN receive and transmit processing, has to be considered as well. The total round trip delay can hence be given as

$$T_{\text{lat}} = T_{\text{FH}} + T_{\text{RAN}} \tag{2.6}$$

$$= T_{\text{proc,FH}} + T_{\text{prop,FH}} + T_{\text{proc,RAN}} \tag{2.7}$$

$$= 2 \cdot \left(N_{\text{node}} \cdot T_{\text{proc,node}} + T_{\text{proc,RU+CU}} + \frac{d_{\text{RU,CU}}}{v_{\text{light,fiber}}} + T_{\text{proc,RAN}} \right), \tag{2.8}$$

with the parameters being explained in Table 2.2. The resulting round trip delays are illustrated in Fig. 2.10 compared to channel coherence times to the three systems of Table 2.1 with two different UE speeds of 3 km/h and 250 km/h. The figures gives both the latencies when including the RAN processing time $t_{\text{proc,RAN}}$ and when not. This is intended to highlight an important point: even when including up to 10 FH nodes, the overall latency is dominated by the RAN processing. While the latency is below the channel coherence time of the two lower carrier systems at low UE mobility, it is far too high for the 70 GHz carrier or higher UE speeds. The conclusion to be drawn from this is, that either RAN processing time needs to be dramatically reduced, or that centralized precoding – i.e. splits A to C – cannot be employed for higher mmWave carriers based on the channel coherence time to be observed.

2.2.4 Time Synchronization

In addition to the total latency, CPRI defines a time synchronization requirement in the form of a delay accuracy, i.e. the delay itself not only needs to be very short, it also needs to be measured precisely. This is motivated by LTE's requirement on aligning samples of different antennas for MIMO diversity transmission. For this, CPRI defines a round trip delay accuracy of 16 ns, which corresponds

Tab. 2.2.: Parameters for FH latency estimation.

Parameter	Symbol	unit	Value	Reference
Number of FH nodes	N_{node}	-	0 ... 10	assumed
Processing time in FH nodes	$T_{\text{proc,node}}$	μs	1	moderate estimation of switch delay [BRW+15]
FH processing time in RU and CU	$T_{\text{proc,RU+CU}}$	μs	5	0.5 round trip time allowed for CPRI [Com]
Distance between RU and CU	$d_{\text{RU,CU}}$	m	100 ... 1000k	assumed
Speed of light on fiber	$v_{\text{light,fiber}}$	m/s	$2 \cdot 10^8$	reduced speed in fiber medium [PWP+13]
RAN processing time	$T_{\text{proc,RAN}}$	ms	2.7	see [SS]

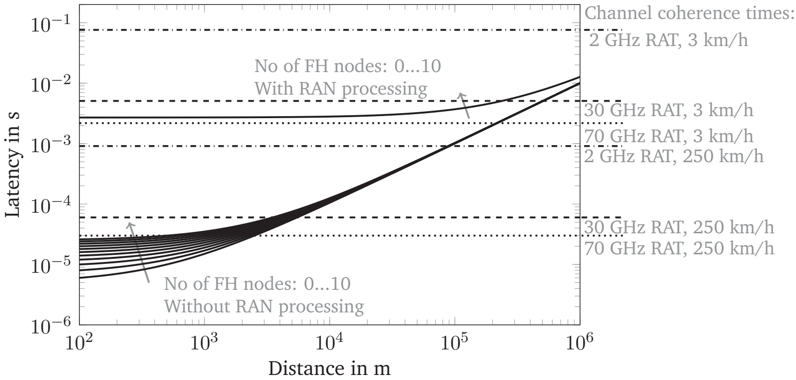

Fig. 2.10.: FH round-trip latencies with and without RAN processing (solid lines), compared to channel coherence times (dashed lines).

to half a sample duration in a 20 MHz LTE system. Accordingly, delay accuracy ΔT_{lat} for the RATs in Table 2.1 can be derived as

$$\Delta T_{\text{lat}} = \frac{1}{2 \cdot f_S} \tag{2.9}$$

$$= 3.33 \text{ ns for 2 GHz RAT} \tag{2.10}$$

$$= 1.33 \text{ ns for 30 GHz RAT} \tag{2.11}$$

$$= 0.67 \text{ ns for 70 GHz RAT.} \tag{2.12}$$

As can be seen, the higher sampling frequencies needed to support the higher bandwidths have a direct impact on the required synchronization. A sub-ns synchronization, as required for the mmWave RATs, will be very challenging to achieve, especially in packet-based transport systems. This aspect will be further discussed in Sec. 2.3.

However, the alignment of the I/Q streams of different antennas or RUs for LTE is only motivated by the limited number of available precoding weights [Eri07], which could be re-evaluated for future networks. A certain delay spread from multipath propagation is expected in any mobile network, and is hence compensated with a cyclic prefix and an equalizer. Accordingly, [BHR13] and references therein give a synchronization requirement of between 3 μs and 10 μs, depending on the cell radius (which in turn has an impact on the maximum delay spread). For mmWave carriers, e.g., [SR15] gives a maximum delay spread on the order of 100 ns, which is still more than 2 orders of magnitude larger than the values derived above. Even when the transport network cannot take up all of the delay spread budget, a synchronization requirement of around 10 ns seems more reasonable.

2.2.5 Frequency Synchronization

In order to align subcarriers properly, the clocks between RUs and CUs need to run at the same frequency and with low variation. For this CPRI defines a maximum clock jitter of 2 part per billion (ppb). It can be assumed that similar requirements will apply to future RATs as well. However, since the given requirement is relative to the clock/sampling frequency, the absolute jitter requirements will be larger for higher bandwidth:

$$\begin{aligned}
\Delta f_{\mathrm{S}} &= f_{\mathrm{S}} \cdot \frac{2}{10^9} && (2.13)\\
&= 0.3 \text{ Hz for 2 GHz RAT} && (2.14)\\
&= 0.75 \text{ Hz for 30 GHz RAT} && (2.15)\\
&= 1.5 \text{ Hz for 70 GHz RAT.} && (2.16)
\end{aligned}$$

Nevertheless, the clock distribution can be expected to be challenging; however, a further analysis is out of scope for this thesis.

2.2.6 Reliability

To guarantee reliable transmission of both user and control data, CPRI requires a bit error rate of less than 10^{-12}. This reliability is commonly achieved only in wired connections, e.g. it is the requirement for 10 Gigabit Ethernet [IEE10, Clause 4.1.2.h]. However, such an error rate is very challenging to achieve in a wireless environment without additional error protection measures such as FECs and HARQ. How to deal with unreliable FH is a key concept of this thesis, and it is hence discussed in detail in Sec. 3.3.

2.3 Enablers for Future Transport Networks

Fiber versus Wireless Fronthaul

Currently, point-to-point fiber links are the technology predominately utilized for FH. It offers high data rates and high reliability. However, the dedicated links are inflexible and deployment is slow due to the involved civil engineering work. In addition, right-of-way has to be acquired to deploy new fiber. Alternatively, fiber capacity can be leased from third parties. However, both options are cost-intensive, making fiber deployments one of the most expensive transport solutions [Rea15].

Alternative solutions can be found in the wireless domain, ranging across different bands. Below Sub-6 GHz, LTE's in-band or out-of-band backhauling [3GP11b] and W-LAN [IEE12] are available. While sub-6 GHz technology is a mature technology and low cost due to the dual use in access links, it shares the limitations of the access technology in terms of range (few hundred meters) and capacity (few Gbps). Especially the low capacity effectively removes the option of utilizing sub-6 GHz bands for lower functional splits.

Microwave links, with frequencies approximately up to 30 GHz are currently already utilized for long-range BH connections. Due to their moderate free-space attenuation they can achieve ranges of several tens of kilometers [HE11]. However, due to their limited bandwidth their data rates are commonly below 1 Gbps, and hence they can barely be employed for lower functional splits in 4G, and will consequently be even less suitable for 5G networks. In addition, most microwave frequencies are regulated, meaning that a license needs to be acquired and payed for, leading to an increased deployment time and operational expenditure (OPEX).

Millimeter wave frequencies, ranging from approximately 30 GHz to 300 GHz, are currently a widely discussed technology for application in 5G access links [RSM+13]. However, they have already been in use for the transport network for several years. Several products that are compliant with CPRI are on the market, achieving ranges of several kilometers, e.g. [Lig]. The main advantage of mmWave frequencies is the high available bandwidth of (depending on the geographical region) up to 7 GHz [Wel09]. With this, data rates of multiple Gbps can be achieved, being suitable even for lower functional splits. In addition, several bands between 50 and 100 GHz are currently either unlicensed or 'lightly licensed', meaning that licenses can be acquired fast and at a low cost. This makes them more cost-efficient and faster to deploy than the traditionally utilized microwave technology or dedicated fiber [Rea15]. It can hence be seen as the most promising alternative to fiber on the last mile.

Challenges of Millimeter Wave Fronthaul

On the downside, the increased frequency of mmWave links compared to microwave results in a higher free-space pathloss. On the other hand, this can be countered by utilizing high-gain antennas, which can be built with smaller form factors. In addition, mmWaves suffer from additional atmospheric attenuation both from oxygen [ITU13a] and precipitation [ITU12]. This not only reduces the maximum achievable range, but also means that mmWave links have to deal with varying channel conditions.

As mmWave FH is such a promising technology but suffers from reliability problems, this thesis investigates unreliable FH and the methods discussed in Chs. 3, 4, and 5 aim to improve C-RAN networks utilizing mmWave links. None withstanding, the concepts introduced can be applied to other technologies sharing similar characteristics as well.

The small antenna element size also enables the application of steerable antenna arrays, similar to the massive MIMO technologies discussed for the access [LETM14; SRH+14]. In the transport, the degree of freedom offered by the arrays would be utilized to steer the beam, thereby enabling dynamic point-to-multi-point connections, adding flexibility that cannot be achieved by fiber links.

Packet-Based Fronthaul

A further factor limiting the flexibility of currently used fiber technology is the utilization of circuit-switched point-to-point links and a proprietary (although publicly available) CPRI protocol. As a result, FH hardware is currently highly specialized and the detailed implementation can even differ from vendor to vendor: this, along with the high required capacities can be seen as one of the reasons for the high costs of FH networks. Packet-switched networks based on Ethernet are currently discussed as an alternative [lIYH+15; HY+15; 5GX16] that has in addition been already widely adopted for BH links. Accordingly, the introduction of packet-based technologies in the FH could lead to a converged FH/BH network, utilizing the same hardware and infrastructure to reduce costs by economics of scale and to simplify management. The introduction of additional functional splits further increases the necessity of a converged solutions, as otherwise additional interfaces and protocols would need to be defined, thereby further increasing complexity.

The main challenges of Ethernet-based transport are twofold: First, a common frame format is required that can be utilized for traffic varying by orders of magnitude in data rate, while requiring different latencies to be met according to the supported application. Second, tight synchronization is not natively supported by Ethernet. As discussed in Sec. 2.2, alignment in time and frequency is required for the lower functional splits. Ethernet, however, is natively asynchronous. Hence, additional technologies need to be applied to make Ethernet applicable to FH. For frequency synchronization, synchronous Ethernet (SyncE) [FGJ+08; ITU13b] is considered as a possible solution. It adds phase locked-loops to network equipment, thereby allowing for a precise recovery of the line clock. As this needs to be supported in hardware, this however requires updates in the equipment already in use. For time alignment, the Precise Timing Protocol (PTP) [IEE08] is currently investigated. PTP utilizes time stamps to estimate transmission delays and align nodes in time. However, this can only operate properly if the delays do not vary considerably due to queueing. So-called transparent clocks can be utilized to estimate wait times in queues, thereby mitigating delay variance. PTP has already been applied in the White Rabbit Project [RL15], where it reportedly achieves sub-ns accuracies. This degree of precision would also be required for centralized mmWave RATs as discussed in Sec.2.2.4, making PTP a promising candidate for packet-switched transport networks. However, here also more advanced equipment will be required, supporting transparent clocks, as well as precise time stamping of packets. These technologies could be applied for any underlying technology such as fiber or mmWave. However, a common packet-based protocol would greatly simplify the interface between the different technologies.

2.4 Chapter Summary

Cloud-RANs are a new concept for 4G and will play a prominent role in 5G networks. However, while introducing many benefits, the centralization also relies on a FH network with very challenging requirements. It is hence of utmost importance to already consider what the impact of 5G RATs will be on the FH network. This chapter showed the way forward from conventional decentralized networks, over centralization to true cloud-based RANs. With the introduction of mmWave RATs, the transport data rates can be expected to increase into the hundreds of Gbps and requiring sub-ns latencies. Different functional splits can be used to reduce the FH requirements in terms of data rate and latency, and especially load-dependent splits can reduce aggregated traffic demand via statistical multiplexing.

The introduction of functional splits calls for a convergence of FH and BH in order to re-use equipment and infrastructure and simplify management. This unified transport network should be packet-based in order to accommodate traffic flows varying in data rate and latency requirements. Several physical link technologies, both wired and wireless, are available for implementation, with especially mmWave links providing a cost-efficient and high-bandwidth solution on the last mile of transport. However, these mmWave links suffer from a reduced reliability as compared to wired links. As reliability is also an important requirement for FH networks and the RAN in general, the rest of this thesis is dedicated to improving FH links utilizing mmWave technology.