# A Mechanism
# to Elastically Utilize Compute Resources
# in Mobile Communication Basestations

Von der Fakultät für Informatik, Elektrotechnik und Informationstechnik
der Universität Stuttgart zur Erlangung der Würde
eines Doktor-Ingenieurs (Dr.-Ing.) genehmigte Abhandlung

vorgelegt von

Thomas Werthmann

geb. in Lennestadt

| | |
|---|---|
| Hauptberichter: | Prof. Dr.-Ing. Andreas Kirstädter |
| Mitberichter: | Prof. Dr.-Ing. Andreas Timm-Giel |
| | |
| Tag der Einreichung: | 10. November 2017 |
| Tag der mündlichen Prüfung: | 27. Juli 2020 |

Institut für Kommunikationsnetze und Rechnersysteme
der Universität Stuttgart
2020

# Abstract

Cellular mobile communication networks have to serve continuously increasing demands in throughput. These demands force to use the limited spectrum more efficiently. One way to achieve this is the application of sophisticated signal processing techniques, e. g. multi-carrier modulation and transmission via multiple antennas (multiple-input multiple-output, MIMO). These techniques generate significant processing effort at the base stations of the network and at the mobile devices. This thesis deals with the processing effort arising in the base stations.

The processing effort is influenced by the radio channel and the transmitted data traffic. It thus varies widely. Dimensioning the processing resources for theoretical peak load is inefficient, because the difference between typical load and peak load is large. However, when processing resources cannot be guaranteed to be sufficient in all situations, the system has to be able to cope with resource shortages. It has to adapt the complexity of the computations, so that the available processing resources are utilized but not overloaded. This ability is here termed elasticity.

This thesis presents a mechanism that allows a base station to elastically utilize the available processing resources. Thereto, first the relationship of network performance and compute resource requirements is formulated as optimization problem. Based on this problem, studies are conducted to identify the components suitable to realize elasticity. Subsequently, a mechanism is designed using the findings from these studies. This mechanism achieves the objective to make the physical layer computations of a base station elastic. Finally, evaluations in two scenarios show the effectiveness of the proposed mechanism.

In principle, the concepts discussed in this thesis apply to different mobile communication standards. However, the presented mechanism is designed to integrate well with the components of a *Long Term Evolution* (LTE) system. LTE scenarios are also used for the evaluations. To provide background for the further discussions, chapter 2 gives an overview over the relevant aspects of the LTE standard. It introduces the LTE standard itself and the system's architecture. Subsequently, the relevant components of the physical layer are explained. These include multi-carrier modulation, coding, and MIMO. Configuring these mechanisms allows to trade the performance of a communication link to reduce computational complexity. This resembles an approach to realize elasticity.

Besides adapting the physical layer configuration, the LTE standard also allows to allocate radio resources dynamically. Radio resources are either allocated to different mobile terminals or left unused. This directly influences the network performance as well as the accruing compute effort. Chapter 3 first discusses various objectives for resource allocation. It then gives an overview over optimization problems and heuristics from literature, which can be applied to solve this task.

When using the same radio spectrum, adjacent cells may interfere with each other's transmissions. By lowering radio resource utilization, interference as well as processing effort can be reduced. As third part of the background information, chapter 4 provides an overview over aspects of multi cell operation. This covers interference, concepts for coordination of cells, and their formulation as optimization problems.

Chapters 5 and 6 cover the main contribution of this thesis. Chapter 5 starts by stating the motivation of the thesis, the research questions, and the contributions. Subsequently, an overview over related work in the topics signal processing, real-time scheduling, and compute resource management is provided. Also, this chapter defines the system model. This is used as foundation for the optimizations and simulations. Besides other aspects, it models interference, physical layer configuration, processing effort, and network performance.

Subsequently, an optimization problem is formulated. This allows the interrelations of relevant effects to be studied without prematurely restricting to heuristical approaches. This optimization problem describes the effect of MIMO mode selection, resource allocation, and interference on the processing requirements and on the network performance. Studying the solutions to this problem validates that it is possible to efficiently cope with limited processing resources. Using a set of related problems, it is also assessed whether a simple method is sufficient to realize the elasticity efficiently. Here, switching between MIMO modes shows to be the most promising approach. The findings from these studies serve as design guidelines for the mechanism described in the following step.

The proposed system consists of a MIMO mode selection heuristic, a mechanism predicting the manageable processing complexity, and a fallback mechanism. The MIMO mode selection balances processing complexity and network performance. This heuristic is executed independently for each mobile terminal. It can thus be distributed over multiple processing units. The operating point of this heuristic is set by predicting the manageable complexity. This prediction is based on the previous time, thus assuming a temporal correlation of the load. The fallback mechanism ensures stable operation in case the prediction is inaccurate. This helps to cope with fluctuations in the load, but also with external disturbances. The proposed system is designed carefully to allow the integration into LTE base stations.

The performance of the proposed system is evaluated thoroughly in chapter 6. Thereto, two scenarios are applied. First, the performance of the proposed system is compared to that of an optimization problem and that of a baseline heuristic in a simplified scenario. These studies show that the proposed approach achieves nearly optimal performance. The second scenario resembles a larger network with dynamic data traffic. Here, compared to the baseline heuristic, the proposed system maintains high network performance even with limited compute resources. Thus, the simple prediction mechanism applied here is sufficient to cope with dynamic load.

Summarizing, the evaluations validate that the proposed approach provides elasticity with high efficiency. This means that the system allows to serve a mobile network with limited compute resources. Beyond that, the network performance is not significantly impacted when compute resources become scarce, given that a minimum amount is available.

Concluding, the proposed system can be implemented to elastically utilize the available processing resources. Instead of strictly requiring sufficient processing capacity to complete all physical layer operations in time, it automatically adapts the complexity of these operations to the available capacity. In doing so, it achieves high efficiency, so that the impact on the network performance is minimized. Thereby, the system allows to dimension compute resources not for a theoretical peak load, but such that they are utilized to capacity. In addition, it lifts the requirement of exact planning and reservation of compute capacity. This facilitates deploying components from standard *information technology* (IT) systems. Overall, the proposed system allows network providers to dimension compute resources economically and thereby increase the cost-efficiency of their network.

# Kurzfassung

**Ein Verfahren zur elastischen Nutzung von Rechenressourcen in Mobilfunk-Basisstationen**

Zelluläre Mobilfunknetze müssen einen ständig wachsenden Datendurchsatz bewältigen. Diese Anforderung zwingt dazu, das limitierte Spektrum effizient zu nutzen. Dies kann durch Anwendung ausgefeilter Techniken der Signalverarbeitung, z. B. Mehrträger-Modulation und Übertragung über mehrere Antennen (multiple-input multiple-output, MIMO), erreicht werden. Diese Techniken erzeugen signifikanten Rechenaufwand in den Basisstationen des Netzes und in den mobilen Geräten. Diese Arbeit befasst sich mit dem Rechenaufwand, der in den Basisstationen anfällt.

Der Rechenaufwand wird vom Funkkanal und vom übertragenen Datenverkehr beeinflusst und schwankt daher stark. Eine Dimensionierung der Rechenressourcen für die theoretische Spitzenlast ist ineffizient, denn der Unterschied zwischen typischer Auslastung und Spitzenlast ist groß. Wenn jedoch die Rechenressourcen so dimensioniert sind, dass sie nicht in allen Situationen ausreichen, muss das System mit Ressourcenengpässen umgehen können. Es muss die Komplexität der Berechnungen anpassen, so dass die verfügbaren Rechenressourcen ausgelastet, aber nicht überlastet werden. Diese Fähigkeit wird hier als Elastizität bezeichnet.

Diese Arbeit präsentiert ein System, welches einer Basisstation ermöglicht, die vorhandenen Rechenressourcen elastisch auszunutzen. Dazu wird zunächst der Zusammenhang zwischen Netzwerk-Leistung und benötigten Rechenressourcen als Optimierungsproblem formuliert. Studien dieses Problems dienen dazu, diejenigen Komponenten zu identifizieren, die sich für die Realisierung der Elastizität eignen. Auf Basis der Ergebnisse wird dann eine Heuristik entworfen. Diese erreicht das Ziel, die Signalverarbeitung einer Mobilfunk-Basisstation elastisch zu machen. Abschließend wird in zwei Szenarien die Effektivität des vorgeschlagenen Systems bewertet.

Die in dieser Arbeit diskutierten Konzepte sind prinzipiell für verschiedene Standards zur Mobilfunk-Kommunikation anwendbar. Die vorgestellte Heuristik ist jedoch so entworfen, dass sie sich gut in die Komponenten eines *Long Term Evolution* (LTE) Systems integrieren lässt. Außerdem werden LTE Szenarien für die Bewertungen verwendet. Um Hintergrundinformationen für die weiteren Diskussionen zu liefern, gibt Kapitel 2 einen Überblick über die relevanten Aspekte des LTE Standards. Das Kapitel führt zunächst in den LTE Standard selbst und die Architektur des LTE Systems ein. Anschließend werden die relevanten Komponenten der physikalischen Schicht erklärt. Dies umfasst unter anderem die Mehrträger-Modulation, Kodierung und MIMO. Die Konfiguration dieser Mechanismen erlaubt es, zwischen der Leistung einer Kommunikationsverbindung und dem anfallenden Rechenaufwand abzuwägen. Dies ist eine Möglichkeit, Elastizität zu realisieren.

Neben der Anpassung der Konfiguration der physikalischen Schicht erlaubt es der LTE Standard auch, Kanal-Ressourcen dynamisch zuzuteilen. Diese Ressourcen werden entweder verschiedenen Endgeräten zugeteilt oder ungenutzt gelassen. Dies beeinflusst direkt die Leistung des Netzes und den anfallenden Rechenaufwand. Kapitel 3 diskutiert zunächst unterschiedliche Zielsetzungen der Ressourcenzuteilung. Es gibt dann einen Überblick über Optimierungsprobleme und Heuristiken aus der Literatur, die angewendet werden können, um diese Aufgabe zu lösen.

Wenn benachbarte Zellen dasselbe Spektrum nutzen, können sie ihre Übertragungen gegenseitig durch Interferenz stören. Werden weniger Kanal-Ressourcen zugeteilt, reduziert das sowohl die Interferenz als auch den Rechenaufwand. Als dritter Teil der Hintergrundinformationen gibt daher Kapitel 4 einen Überblick über die Aspekte des Betriebs mehrerer Zellen. Die behandelten Themen umfassen Interferenz, Konzepte zur Koordination von Zellen und deren Formulierung als Optimierungsproblem.

Kapitel 5 und 6 stellen den Hauptbeitrag dieser Arbeit dar. Kapitel 5 beginnt damit, die Motivation der Arbeit zu erläutern und die Forschungsfragen und die Beiträge zu benennen. Im Anschluss wird diese Arbeit in den Kontext der vorhergehenden Forschung in den Bereichen Signalverarbeitung, Echtzeit-Scheduling und Verwaltung von Rechenressourcen eingeordnet. Daraufhin wird das Systemmodell definiert. Dieses wird im weiteren Verlauf als Basis für die Optimierungsprobleme und Simulationsstudien verwendet. Es modelliert unter anderem Interferenz, Konfiguration der physikalischen Schicht, Rechenaufwand und Netzwerk-Leistung.

Im Anschluss wird ein Optimierungsproblem formuliert, um den Zusammenhang relevanter Effekte ohne vorzeitige Beschränkung auf heuristische Ansätze zu untersuchen. Dieses Optimierungsproblem beschreibt den Einfluss von Interferenz, Zuteilung der Kanal-Ressourcen und Auswahl des MIMO Modus auf den anfallenden Rechenaufwand und die Leistung des Netzes. Die Auswertung der Lösungen dieses Problems bestätigt, dass es möglich ist, auf effiziente Art mit limitierten Rechenressourcen umzugehen. Mit verwandten Problemen wird dann untersucht, ob ein einfaches Vorgehen für die effiziente Realisierung von Elastizität ausreichend ist. Hier zeigt sich der Wechsel zwischen MIMO Modi als vielversprechendster Ansatz. Die Erkenntnisse aus diesen Studien dienen als Design-Vorgaben für die im Folgenden beschriebene Heuristik.

Das vorgeschlagene System besteht aus einer Heuristik zur Auswahl des MIMO Modus, einem Mechanismus zur Vorhersage der zu bewältigenden Rechen-Komplexität und einem Rückfall-Mechanismus. Die Auswahl des MIMO Modus definiert den Kompromiss zwischen Rechen-Komplexität und Netzwerk-Leistung. Diese Heuristik wird für jedes Endgerät unabhängig ausgeführt, so dass eine Parallelisierung in mehreren Recheneinheiten möglich ist. Der Betriebspunkt dieser Heuristik wird durch die Vorhersage der zu bewältigenden Rechen-Komplexität eingestellt. Diese Vorhersage basiert auf der vergangenen Zeit, setzt also eine zeitliche Korrelation der Last voraus. Der Rückfall-Mechanismus garantiert ein stabiles Verhalten des Systems auch dann, wenn die Vorhersage ungenau ist. Dadurch kann das System sowohl mit Schwankungen in der Last als auch mit Störungen von außen umgehen. Das System wurde so entworfen, dass es in LTE Basisstationen integriert werden kann.

Die Leistung des vorgeschlagenen Systems wird in Kapitel 6 ausführlich bewertet. Dazu werden zwei Szenarien verwendet. Zuerst wird in einem einfachen Szenario die Leistung des vorgeschlagenen Systems mit der eines Optimierungsproblems und einer Basis-Heuristik verglichen. Diese Studien zeigen, dass das vorgeschlagene System nahezu optimale Leistung erreicht. Das zweite Szenario modelliert ein größeres Netzwerk mit dynamischem Datenverkehr. Im Vergleich zur Basis-Heuristik zeigt das vorgeschlagene System hier, dass hohe Leistung auch mit eingeschränkten Rechenressourcen aufrechterhalten werden kann. Der hier verwendete einfache Vorhersage-Mechanismus ist also ausreichend, um mit dynamischer Last umzugehen.

Zusammenfassend weisen die Studien nach, dass das vorgeschlagene System die Elastizität des Mobilfunksystems mit hoher Effizienz realisiert. Das System ermöglicht, ein Mobilfunk-Netz grundsätzlich auch mit eingeschränkten Rechenressourcen zu betreiben. Darüber hinaus wird, wenn die Rechenressourcen verknappt werden, die Leistung des Netzes nicht signifikant beeinträchtigt, solange dabei ein Minimalwert nicht unterschritten wird.

Das vorgeschlagene System kann implementiert werden, um die verfügbaren Rechenressourcen elastisch auszulasten. Anstatt vorauszusetzen, dass genügend Rechenkapazität verfügbar ist, um immer alle Berechnungen der physikalischen Schicht rechtzeitig abzuschließen, passt es automatisch die Komplexität dieser Berechnungen an die verfügbare Rechenkapazität an. Dabei erreicht es eine hohe Effizienz, so dass die Beeinträchtigung der Netzwerk-Leistung minimiert wird. Rechenressourcen müssen dadurch nicht für die theoretische Spitzenlast dimensioniert werden. Sie können stattdessen so ausgelegt werden, dass sie optimal ausgelastet werden. Zusätzlich hebt das System die Anforderung auf, Rechenleistung exakt zu planen und zu Rechenressourcen zu reservieren. Dies erleichtert den Einsatz von Komponenten aus klassischen *Information Technology* (IT) Systemen. Insgesamt erlaubt das vorgeschlagene System Betreibern von Mobilfunknetzen, Rechenleistung ökonomisch zu dimensionieren und damit die Kosteneffizienz des Netzes zu steigern.

# Contents

# List of Figures

# List of Tables

# Abbreviations

| | | |
|---|---|---|
| **3GPP** | Third Generation Partnership Project | 1 |
| **ABS** | almost blank subframe | 75 |
| **AC** | admission control | 49 |
| **ADC** | analog-to-digital converter | 2 |
| **AMC** | adaptive modulation and coding | 23 |
| **AMPS** | Advanced Mobile Phone System | 8 |
| **ARQ** | automatic repeat request | 23 |
| **ASIC** | application-specific integrated circuit | 16 |
| **AWGN** | additive white Gaussian noise | 19 |
| **BBU** | baseband unit | 2 |
| **BCCH** | Broadcast Control Channel | 14 |
| **BCH** | Broadcast Channel | 15 |
| **BET** | Blind Equal Throughput | 61 |
| **BLER** | block error rate | 98 |
| **BS** | base station | 1 |
| **CA** | carrier aggregation | 40 |
| **CCCH** | Common Control Channel | 14 |
| **CCE** | control channel element | 46 |
| **CDF** | cumulative distribution function | 55 |
| **CoMP** | coordinated multi-point | 8 |
| **CP** | cyclic prefix | 20 |
| **CPRI** | Common Public Radio Interface | 16 |
| **CPU** | central processing unit | 167 |
| **CQI** | channel quality indicator | 35 |
| **C-RAN** | cloud RAN | |
| **CRC** | cyclic redundancy check | 22 |
| **CRS** | Cell-Specific Reference Signal | 29 |

| | | |
|---|---|---|
| **CSI** | channel state information | 25 |
| **CSI-RS** | CSI Reference Signal | 35 |
| **CSMA/CA** | Carrier Sense Multiple Access / Collision Avoidance | 59 |
| **DAC** | ditigal-to-analog converter | 2 |
| **DCCH** | Dedicated Control Channel | 14 |
| **DCI** | downlink control information | 46 |
| **DFT** | discrete Fourier transform | 20 |
| **DL** | downlink | 3 |
| **DL-SCH** | Downlink Shared Channel | 15 |
| **DM-RS** | Demodulation Reference Signal | 30 |
| **DMT** | discrete multitone transmission | 20 |
| **DPD** | digital pre-distortion | 38 |
| **DP** | double precision | 185 |
| **D-RAN** | distributed RAN | 2 |
| **DRX** | discontinuous reception | 84 |
| **DSL** | digital subscriber line | 13 |
| **DSP** | digital signal processor | 16 |
| **DTCH** | Dedicated Traffic Channel | 14 |
| **ECN** | explicit congestion notification | 50 |
| **eNodeB** | evolved nodeB | 1 |
| **EPC** | evolved packet core | 7 |
| **EPDCCH** | Enhanced PDCCH | 15 |
| **EPS** | evolved packet system | 7 |
| **ETSI** | European Telecommunications Standards Institute | 51 |
| **E-UTRAN** | evolved UMTS terrestrial radio access network | 10 |
| **FDD** | frequency division duplex | 39 |
| **FEC** | forward error correction | 12 |
| **FFT** | fast fourier transform | 20 |
| **FLOP** | floating point operation | 101 |
| **FPGA** | field-programmable gate array | 16 |
| **FSTD** | frequency-switched transmit diversity | 26 |
| **FTP** | file transfer protocol | 99 |
| **GBR** | guaranteed bitrate | 52 |
| **GOPS** | giga operations per second | 101 |
| **GPP** | general purpose processor | 16 |

| | | |
|---|---|---|
| **GSM** | Global System for Mobile Communications, orig. Groupe Spécial Mobile | 1 |
| **HARQ** | hybrid ARQ | 23 |
| **HII** | high interference indication | 75 |
| **HOL** | head of line | 61 |
| **HSPA** | High Speed Packet Access | 8 |
| **HSS** | Home Subscriber Server | 11 |
| **HTML** | hypertext markup language | 51 |
| **HTTP** | hypertext transfer protocol | 99 |
| **IAT** | interarrival time | 87 |
| **ICIC** | inter-cell interference coordination | 68 |
| **IEEE** | Institute of Electrical and Electronics Engineers | 10 |
| **IETF** | Internet Engineering Task Force | 51 |
| **IfCo** | interference coordination | 68 |
| **IMT** | International Mobile Telecommunications | 8 |
| **IP** | Internet Protocol | 1 |
| **ISI** | inter symbol interference | 20 |
| **IT** | information technology | 3 |
| **ITU** | International Telecommunications Union | 8 |
| **LA** | link adaptation | 23 |
| **LoS** | line of sight | 18 |
| **LP** | linear programming | 133 |
| **LTE-A** | LTE-Advanced | 1 |
| **LTE-A Pro** | LTE-Advanced Pro | 1 |
| **LTE** | Long Term Evolution | 1 |
| **MAC** | Medium Access Control | 12 |
| **MBMS** | Multimedia Broadcast Multicast Service | 9 |
| **MBR** | maximum bitrate | 52 |
| **MCCH** | Multicast Control Channel | 14 |
| **MCH** | Multicast Channel | 15 |
| **MCKP** | multiple-choice knapsack problem | 134 |
| **MCS** | modulation and coding scheme | 23 |
| **MIESM** | mutual information effective SINR metric | 97 |
| **MIMO** | multiple-input multiple-output | 1 |
| **ML** | maximum likelyhood | 29 |
| **M-LWDF** | Modified Largest Weighted Delay First | 61 |

| | | |
|---|---|---|
| **MME** | Mobility Management Entity | 11 |
| **MMSE** | minimum mean squared error | 28 |
| **MTCH** | Multicast Traffic Channel | 14 |
| **MU-MIMO** | multi-user MIMO | 27 |
| **NAICS** | network assisted interference cancellation and suppression | 66 |
| **NFV** | network function virtualization | 17 |
| **NGMN** | Next Generation Mobile Networks | 55 |
| **NLoS** | non line of sight | 18 |
| **NMT** | Nordic Mobile Telephone | 8 |
| **NTT** | Nippon Telegraph and Telephone | 8 |
| **OFDMA** | orthogonal frequency division multiple access | 1 |
| **OFDM** | orthogonal frequency division multiplexing | 20 |
| **OI** | overload indication | 75 |
| **OSI** | Open Systems Interconnection | 12 |
| **OS** | operating system | 17 |
| **PAPR** | peak-to-average power ratio | 21 |
| **PBCH** | Physical Broadcast Channel | 15 |
| **PCCH** | Paging Control Channel | 14 |
| **PCFICH** | Physical Control Format Indicator Channel | 15 |
| **PCH** | Paging Channel | 15 |
| **PDCCH** | Physical Downlink Control Channel | 15 |
| **PDCP** | Packet Data Convergence Protocol | 12 |
| **PDF** | probability density function | 100 |
| **PDSCH** | Physical Downlink Shared Channel | 15 |
| **PF** | Proportional Fair | 54 |
| **P-GW** | Packet Data Network Gatewayalso PDN-GW | 11 |
| **PHICH** | Physical HARQ Indicator Channel | 15 |
| **PMCH** | Physical Multicast Channel | 15 |
| **PMI** | precoding matrix indicator | 36 |
| **PRACH** | Physical Random Access Channel | 15 |
| **PRB** | physical resource block | 41 |
| **PSS** | Primary Synchronization Signal | 42 |
| **PUCCH** | Physical Uplink Control Channel | 15 |
| **PUSCH** | Physical Uplink Shared Channel | 15 |
| **QAM** | quadrature amplitue modulation | 20 |

| | | |
|---|---|---|
| **QCI** | QoS class identifier | 52 |
| **QoS** | quality of service | 11 |
| **QPSK** | quadrature phase shift keying | 20 |
| **RACH** | Random Access Channel | 15 |
| **RAN** | radio access network | 2 |
| **RA** | resource allocation | 4 |
| **RBG** | resource block group | 45 |
| **RE** | resource element | 41 |
| **RI** | rank indication | 36 |
| **RLC** | Radio Link Control | 12 |
| **RNTP** | relative narrowband transmit power | 75 |
| **R-PDCCH** | Relay PDCCH | 15 |
| **RRC** | Radio Resource Control | 12 |
| **RRH** | remote radio head | 2 |
| **RS** | reference symbol | 21 |
| **RT** | real-time | 3 |
| **SAE** | System Architecture Evolution | 7 |
| **SC-FDMA** | single carrier frequency division multiple access | 21 |
| **SCM** | Spatial Channel Model | 95 |
| **SDN** | software defined networking | 17 |
| **SDR** | software defined radio | 16 |
| **SDU** | service data unit | 12 |
| **SFBC** | space-frequency block coding | 26 |
| **S-GW** | Serving Gateway | 11 |
| **SIC** | successive interference cancellation | 29 |
| **SINR** | signal to interference and noise ratio | 19 |
| **SISO** | single-input single-output | 24 |
| **SNR** | signal to noise ratio | 19 |
| **SON** | self-organizing networks | 65 |
| **SP** | single precision | 185 |
| **SRS** | Sounding Reference Signal | 34 |
| **SSS** | Secondary Synchronization Signal | 42 |
| **SVD** | singular value decomposition | 27 |
| **TCP** | Transmission Control Protocol | 24 |
| **TDD** | time division duplex | 40 |

| | | |
|---|---|---|
| **TDMA** | time division multiple access | 61 |
| **TETRA** | Terrestrial Trunked Radio | 1 |
| **TM** | transmission mode | 30 |
| **TP** | transmission point | 76 |
| **TTI** | transmission time interval | 40 |
| **UE** | user equipment | 10 |
| **UL-SCH** | Uplink Shared Channel | 15 |
| **UL** | uplink | 9 |
| **UMTS** | Universal Mobile Telecommunications System | 1 |
| **VM** | virtual machine | 80 |
| **VoIP** | voice over IP | 8 |
| **VRB** | virtual resource block | 44 |
| **WLAN** | Wireless Local Area Network | 59 |
| **ZF** | zero-forcing | 28 |

# Mathematical Notation and Symbols

## Letter Style

| | |
|---|---|
| $\mathbf{a}, \mathbf{b}, \mathbf{c}$ | Vectors |
| $\mathbf{A}, \mathbf{B}, \mathbf{C}$ | Matrices |
| $\mathcal{A}, \mathcal{B}, \mathcal{C}$ | Sets |

## Operators

| | |
|---|---|
| $\wedge$ | Logical conjunction (and) |
| $\neg$ | Logical negation (not) |
| $\vee$ | Logical disjunction (or) |
| $()^{-1}$ | Inverse of a matrix |
| $()^{\dagger}$ | Pseudo-inverse of a matrix |
| $()^{\mathsf{H}}$ | Hermitian transpose of a matrix, also known as conjugate transpose (the transposed matrix, where all entries are replaced with their complex conjugate) |
| $\mathbb{E}[\bullet]$ | Expected value |
| $\in$ | Element of a set, e. g. $a \in \mathcal{A}$: $a$ is an element of $\mathcal{A}$ |
| $\{\bullet\}$ | Definition of a set with the given elements, e. g. $\mathcal{A} = \{a\}$ is a set with the single element $a$ |
| $\subset$ | Subset of a set, e. g. $\mathcal{B} \subset \mathcal{A}$: $\mathcal{B}$ is a subset of $\mathcal{A}$ |
| $\|\bullet\|$ | Cardinality of a set, i. e. the number of elements |
| $\cap$ | Intersection of two sets, e. g. $\mathcal{B} \cap \mathcal{A}$: intersection of $\mathcal{A}$ and $\mathcal{B}$ |
| $\mathcal{P}(\bullet)$ | Power set (i. e. the set of all possible subsets), e. g. with $\mathcal{A} = \{1, 2\}$, $\mathcal{P}(\mathcal{A}) = \{\{1, 2\}, \{1\}, \{2\}, \{\}\}$ |

| | |
|---|---|
| $\cup$ | Union of two sets, e. g. $\mathcal{A} \cup \mathcal{B}$: union of $\mathcal{A}$ and $\mathcal{B}$ |
| $\bigcup_{\bullet}$ | Union of multiple sets, e. g. $\bigcup_{i \in \{\mathcal{A},\mathcal{B},\mathcal{C}\}} = \mathcal{A} \cup \mathcal{B} \cup \mathcal{C}$ |
| $\bullet \setminus \bullet$ | Subtraction of sets, e. g. with $\mathcal{A} = \{1, 2\}$, $\mathcal{A} \setminus \{1\} = \{2\}$ |

## Symbols

| | |
|---|---|
| $a$ | Number of transmit antennas (parameter of the processing effort model) |
| $\breve{a}_s$ | Variable denoting the scaling of all resource allocations in subframe $s$ |
| $a_{u,f}$ | Flag describing the allocation of PRB $f$ to UE $u$ in the current subframe |
| $\check{a}_{u,m}$ | Continuous variable denoting to which fraction MIMO mode $m$ is used by UE $u$ |
| $\hat{a}_{u,m}$ | Flag denoting whether MIMO mode $m$ is used by UE $u$ |
| $a_{u,n^{\mathrm{c}}}^{(\infty)}$ | Flag defining whether resource $n^{\mathrm{c}}$ is used to serve UE $u$, given that the compute resources are not restricted |
| $a_{u,n^{\mathrm{c}},m}$ | Flag denoting whether resource $n^{\mathrm{c}}$ is used to serve UE $u$ with MIMO mode $m$ |
| $\tilde{a}_{u,s,m}$ | Flag denoting whether MIMO mode $m$ is used to serve UE $u$ in subframe $s$ |
| $\mathcal{B}$ | Set of all BSs in the system |
| $b$ | System bandwidth (parameter of the processing effort model) |
| $\mathcal{B}_s^{\mathrm{active}}$ | Set of BSs which are active in system state $s$ |
| $b_u^\star$ | Serving BS of UE $u$ |
| $\mathbb{C}$ | Set of complex numbers |
| $c$ | Code rate (parameter of the processing effort model) |
| $\mathrm{cap}(\gamma)$ | Channel capacity as function of the SINR $\gamma$ |
| $d_{\mathrm{f}}$ | Utilization in frequency domain (parameter of the processing effort model) |
| $d_{\mathrm{is}}$ | Inter site distance, distance between adjacent BS sites |
| $d_{\mathrm{t}}$ | Utilization in time domain (parameter of the processing effort model) |
| $e_{\mathrm{min}}$ | Efficiency threshold, MIMO modes with lower efficiency are not used |
| $e_{\mathrm{min}}^{\mathrm{opt}}$ | Global efficiency threshold used as variable in the optimization |
| $e_{\mathrm{min},s}^{\mathrm{opt}}$ | Efficiency threshold for subframe $s$ used as variable in the optimization |
| $e_{\mathrm{step}}$ | Efficiency step size used by the prediction mechanism |

| | |
|---|---|
| $e_{u,m}$ | Efficiency of MIMO mode $m$ for UE $u$ |
| $(e^{(-)}_{u,s,m}, e^{(+)}_{u,s,m}]$ | Efficiency interval in which MIMO mode $m$ is used to serve UE $u$ in subframe $s$ |
| $F$ | Noise figure |
| $\mathcal{F}$ | Set of all frequencies (PRB indices) |
| $\tilde{\gamma}_n$ | Post-equalization SINR on spatial layer $n$ |
| $\gamma_{u,b}$ | Macro-scale attenuation between BS $b$ and UE $u$ |
| $\mathbf{H}$ | Channel matrix in a MIMO system |
| $\mathbf{h}$ | Channel vector in a MIMO system |
| $h_{b,u,n^{\mathrm{p}}}$ | Channel gain from BS $b$ to UE $u$ on power allocation resource $n^{\mathrm{p}}$ |
| $\mathbf{H}_{\mathrm{eff}}$ | Effective channel matrix in a MIMO system, i. e. including precoding |
| $\mathbf{h}_{\mathrm{eff}}$ | Effective channel vector in a MIMO system, i. e. including precoding |
| $\widehat{\mathbf{H}}_{u,b}(t, f)$ | Channel matrix describing time-varying, frequency selective small-scale fading between BS $b$ and UE $u$, exlcuding the macro-scale attenuation $\gamma_{u,b}$ |
| $\mathbf{I}$ | Identity matrix |
| $\mathcal{I}_u$ | Set of interferers of UE $u$ |
| $\mathcal{I}_u^{\mathrm{rel}}$ | Set of relevant interferers of UE $u$ |
| $l$ | Number of spatial streams (parameter of the processing effort model) |
| $\mathcal{M}$ | Set of all MIMO modes |
| $m$ | Modulation bits per symbol (parameter of the processing effort model) |
| $\widetilde{\mathcal{M}}_u^{\mathrm{cand}}$ | Set of candidate MIMO modes considered for UE $u$ |
| $\widetilde{\mathcal{M}}_u$ | Set of MIMO modes that are neither dominated nor LP-dominated for UE $u$ |
| $m_{u,n}^{\mathrm{PF}}$ | Per-PRB RA metric of the PF algorithm, for UE $u$ and PRB $n$ |
| $m^{\star}_{u,n^{\mathrm{g}}}$ | MIMO mode which results in the highest spectral efficiency for UE $u$ on a resource in power allocation group $n^{\mathrm{g}}$ |
| $N_0$ | Noise power spectral density |
| $N_{\mathrm{batches}}$ | Number of simulation batches |
| $N_{\mathrm{cand}}$ | Number of candidate UEs considered by the RA algorithm for each BS and subframe |
| $N_{\mathrm{I}}$ | Number of considered relevant interferers per UE |

$N_{\mathrm{PRB}}$ Number of PRBs per TTI

$N_{\mathrm{Rx}}$ Number of receive antennas of a MIMO system

$N_{\mathrm{Tx}}$ Number of transmit antennas of a MIMO system

$p_{b,n^{\mathrm{p}}}$ Transmit power of BS $b$ on power allocation resource $n^{\mathrm{p}}$

$P_n^{\mathrm{peak}}$ Theoretical peak effort of a system with $n$ PRBs

$p_{\max}$ Transmit power limit per BS $b$ and power allocation resource

$p_{\max}$ Available compute resources as fraction of the theoretical peak compute requirments of the system

$p_{\max}^{\mathrm{abs}}$ Available compute resources (absolute value)

$p_{n^{\mathrm{m}}}$ Compute effort per unit resource size on MIMO resource $n^{\mathrm{m}}$

$p_{\mathrm{off}}$ Compute effort offset used by the prediction mechanism

$p_{\mathrm{total}}^{\mathrm{real}}$ The acual effort invested to encode data for all UEs

$p_u^{\mathrm{real}}$ The actual effort invested to encode the data for UE $u$ (in contrast, $p_{u,m}$ is the effort estimated before data is encoded)

$p_{\mathrm{rem}}$ Remaining processing capacity, given the selection of MIMO modes currently considered by the algorithm

$P_{\mathrm{total}}$ Processing effort caused by the whole system (resulting from the processing effort model)

$P_{\mathrm{total}}$ Total transmit power

$\hat{p}_{u,m}$ Additional compute effort incurred by using MIMO mode $m$ for UE $u$, compared to the MIMO mode with the next lower complexity

$p_{u,m}$ Compute effort per unit resource size when serving UE $u$ with MIMO mode $m$

$\mathbb{R}$ Set of real numbers

$r_u^{\mathrm{act}}(t)$ Rate allocated to UE $u$ by RA at time $t$

$\mathcal{R}^{\mathrm{c}}$ Set of all considered channel resources, i. e. two-dimensional grid of subframes and PRBs

$\mathcal{R}_s^{\mathrm{c}}$ Set of channel resources of subframe $s$ (the PRBs)

$\mathcal{R}_b^{\mathrm{g}}$ Set of groups (subsets) of power allocation resources, grouped by state of relevant interferers

$r^{\max}$ Maximum rate achievable by a single UE

$\mathcal{R}^{\mathrm{m}}$ Set of all MIMO resources, one for each combination of UE in $\mathcal{U}$, power allocation resource in $\mathcal{R}^{\mathrm{p}}$, and MIMO mode in $\mathcal{M}$

| | |
|---|---|
| $\mathcal{R}^{\mathrm{m}}_{n^{\mathrm{s}}_{u,n^{\mathrm{p}}}}$ | Set of MIMO resources belonging to the scheduling resource $n^{\mathrm{s}}_{u,n^{\mathrm{p}}}$ |
| $\mathcal{R}^{\mathrm{m}}_{u,n^{\mathrm{p}}}$ | Set of MIMO resources of UE $u$ on power allocation resource $n^{\mathrm{p}}$, one for each MIMO mode in $\mathcal{M}$ |
| $r^{\mathrm{min}}$ | Minimum rate guaranteed to each UE |
| $r_{n^{\mathrm{m}}}$ | Data rate per unit resource size on MIMO resource $n^{\mathrm{m}}$ |
| $\mathcal{R}^{\mathrm{p}}$ | Set of power allocation resources |
| $\mathbb{R}^{+}$ | Set of positive real numbers |
| $\mathcal{R}^{\mathrm{s}}$ | Set of all scheduling resources, one for each combination of power allocation resource in $\mathcal{R}^{\mathrm{p}}$ and UE in $\mathcal{U}_b$ |
| $\mathcal{R}^{\mathrm{s}}_{b,n^{\mathrm{p}}}$ | Set of scheduling resources of BS $b$ on power allocation resource $n^{\mathrm{p}}$, one for each UE in $\mathcal{U}_b$ |
| $r_u$ | Total rate achieved by UE $u$ |
| $r_{u,n^{\mathrm{p}}}$ | Data rate achieved by UE $u$ on power allocation resource $n^{\mathrm{p}}$ |
| $r_{u,n}(t)$ | Channel capacity for UE $u$ on PRB $n$ at time $t$, precited for RA |
| $\bar{R}_u(t)$ | Previous data rate for UE $u$, evaluated at time $t$ |
| $\mathcal{S}$ | Set of all system states |
| $\mathcal{S}$ | Set of all subframes |
| $s_{\mathrm{max}}$ | Maximum number of resources (PRBs) the sytem can allocate |
| $s_{n^{\mathrm{p}}}$ | Size of a resource, e. g. a power allocation resource $n^{\mathrm{p}}$ or a scheduling resource $n^{\mathrm{s}}$ |
| $s^{\mathrm{start}}_{u}$ | Subframe in which the transmission to UE $u$ starts |
| $\mathcal{T}_b$ | Set of state groups of BS $b$ |
| $t_{\mathrm{batch}}$ | Duration of a simulation batch |
| $t_c$ | Time constant of the low pass filter used by the PF RA algorithm |
| $t^{\mathrm{max}}_{c}$ | Upper bound of the time constant value |
| $t_{c,u}(s)$ | Time constant of the low pass filter used by the PF RA algorithm, here depending on the UE and the subframe $s$ |
| $t_{\mathrm{IAT}}$ | Inter arrival time of the data traffic model |
| $t_{\text{warm-up}}$ | Duration of the warm-up phase of the simulation |
| $\mathcal{U}$ | Set of all UEs in the system |
| $\mathcal{U}_b$ | Set of all UEs served by BS $b$ |

| | |
|---|---|
| $\mathrm{U}_{\mathrm{cap}}(\bullet)$ | Utility as function of data capacity |
| $\mathrm{U}(\bullet)$ | Utility as function of rate |
| $\mathrm{U}_{u}(\bullet)$ | Utility of UE $u$ as function of rate |
| $\upsilon_{u,m}$ | Precalculated utility achieved when UE $u$ uses MIMO mode $m$ |

# 1 Introduction

Over the past decades, commercial wireless networks have developed in multiple aspects. The most prominent of these is the efficient utilization of the radio spectrum. The term *spectral efficiency* denotes the delivered throughput divided by the system bandwidth. Since the start of digital data transmission in cellular wireless systems, it has been increased by more than an order of magnitude. However, for economically efficient operation of such networks, investment and operational costs cannot be neglected.

Efficient utilization of compute hardware contributes to cost-efficient network operation. Multiple mechanisms, algorithms, and realization approaches allow to trade-off between computational complexity and network capacity. This thesis proposes to dynamically adjust these trade-offs to always make best use of the available compute resources. It presents a mechanism which achieves this by adapting parameters of the transmission.

## 1.1 Cellular Wireless Communication

In a cellular wireless network, a large service area is split into multiple cells. In each cell, multiple optionally mobile users are served by a single stationary *base station* (BS). This BS manages the resources of the cell, which are shared by the assigned users. Cells operate independently of each other. Their spatial separation allows re-using the radio spectrum. Prominent examples for cellular networks are *Global System for Mobile Communications* (GSM), *Universal Mobile Telecommunications System* (UMTS), and *Long Term Evolution* (LTE) for commercial services to consumers, and *Terrestrial Trunked Radio* (TETRA) for governmental use.

The currently newest, widely deployed commercial cellular network is LTE with its successors *LTE-Advanced* (LTE-A) and *LTE-Advanced Pro* (LTE-A Pro). These are standards developed by the *Third Generation Partnership Project* (3GPP). In LTE, the BSs are termed *evolved nodeB* (eNodeB). They are typically located close to the cell tower, where the corresponding antennas are mounted. An eNodeB contains analog components, e. g. filters and power amplifiers, and also handles all digital processing for the wireless communication. This includes the protocol stack up to the *Internet Protocol* (IP) layer.

LTE employs *orthogonal frequency division multiple access* (OFDMA) to serve multiple users in parallel on different radio frequencies. The system adapts to the characteristics of the individual radio channels to get the most out of available radio spectrum. This comprises the application of different modulation schemes and *multiple-input multiple-output* (MIMO) modes, which use multiple antennas at the BSs and users' terminals to increase the network's throughput.

C-RAN stands for *centralized radio access network* (RAN) or *cloud* RAN. It is an alternative approach for the implementation of the eNodeBs. Here, each eNodeB is split into the *remote radio head* (RRH) and the *baseband unit* (BBU). A RRH contains *analog-to-digital converters* (ADCs), *ditigal-to-analog converters* (DACs), and all analog components for transmission and reception of signals. The BBU performs digital signal processing and higher layer operations. Multiple BBUs are co-located in a central office. RRHs and BBUs exchange data over high speed links. In demarcation to C-RAN, the classical implementation consisting of distributed self-contained eNodeBs is termed *distributed RAN* (D-RAN).

The C-RAN architecture reduces maintenance cost and allows for easier cooperation between multiple cells. Co-located BBUs can share the resources of a common hardware pool. Optionally, virtualization techniques can be used to make the association of compute tasks to hardware flexible. The objective is to balance load, to quickly recover from broken hardware, and to be able to take hardware offline for maintenance.

## 1.2 Compute Requirements of Wireless Communication Systems

Today's communication systems invest considerable computational effort for signal processing, encoding/decoding of information, protocol processing, and resource management. This is carried out by a combination of hardware and software. Both are traditionally designed and assembled for a single communication protocol.

In such a design, the split between hardware and software is debatable.[1] On one hand, hardware designed for a single task is typically more efficient than generic hardware, i. e. it performs the task more quickly and with lower power consumption. On the other hand, engineering of software implementations is often considered to be easier than that of the respective hardware implementations. In addition, software development cycles are shorter, and software can be updated in already deployed products to remedy problems and add new features. For the implementation of network infrastructure, there is currently a trend to avoid specialized hardware.

Each user terminal has a varying demand for communication. The BS allocates bandwidth to satisfy these demands. In doing so, it adapts parameters of the transmissions to the respective radio channels. The changing utilization of bandwidth and the varying parameters influence the compute effort. Thus, the compute effort accruing for the operation of an eNodeB fluctuates.

The compute effort of a single cell is limited by the most complex communication parameters and the available spectrum resources. If multiple BBUs share a pool of compute resources, their compute effort is multiplexed. Thereby, the compute load of such a pool behaves more smoothly than that of a single BBU. The spatial distribution of the cells belonging to the pool results in averaging also over the spatial inhomogeneity of the network load.

[1] Although a software-based installation is often assumed to be present in C-RAN setups, this question can in principle be seen as orthogonal to the selection of a C-RAN or a D-RAN architecture.

## 1.3 Elastic Utilization of Compute Resources

For many signal processing systems, the computational complexity can be planned in advance. The resources are then dimensioned to cope with the peak load. This guarantees that the resources are always sufficient. Consequently, the running system does not need to actively manage resource utilization. However, if the compute load of the system is variable, this results in an inefficient utilization of the resources. Besides, if the availability of the compute resources is impacted by unforeseen events, the system cannot operate as desired.

In *information technology* (IT) systems without *real-time* (RT) requirements, increasing the load or reducing the processing resources result in delayed responses of the system. This is not allowed in a RT system like the eNodeB in an LTE network. Thus, to cope with resource limitations, RT systems have to be designed to actively adapt to the availability of compute resources. This capability is here termed *elasticity*.

Adapting to scarce compute capabilities implicates reduced system performance. However, this is required to maintain system operation if resources are not sufficient otherwise. If implemented well, the price to save some compute effort is small. For example, a system can switch to a mode of operation with slightly impacted performance, but much lower computational complexity.

Deploying an elastic signal processing system in the BBUs of a C-RAN brings multiple benefits. First, it allows to dimension the compute hardware to meet the typical resource requirements instead of the peak requirements. For systems which support complex, but rarely used transmission modes, this can save a significant fraction of compute hardware. Second, it enables the operator to save energy by taking down hardware resources during low load periods. When the load increases, the elasticity of the system bridges the time it takes to bring resources up again. Third, an elastic system can cope with planned or unplanned outage of resources. In case resources are offline, either caused by hardware failures or for maintenance, the system performance degrades gracefully. The network operator can therefore avoid installing spare resources.

Summarizing, using elastic implementations makes the operation of a pool of BBUs more similar to the operation of IT software. Instead of deploying specialized RT capable operating systems and virtualization software, standard implementations can be used. Instead of generously over-dimensioning resources to avoid shortages, virtualization and elasticity can work together to allow high resource utilization and thereby efficient network operation.

## 1.4 Research Questions and Contributions

The main objective of this thesis is to enable a group of pooled LTE BBUs to elastically utilize their shared compute resources. This is split into three major questions and tasks.

Multiple components of an LTE system allow to trade system performance for compute effort reduction. Thus, the first question is to identify those components where this can be performed efficiently. Thereby, the focus here lies on the signal processing for *downlink* (DL) operation. To avoid premature restriction to simplifying heuristics, different approaches to reduce the processing effort are formulated as optimization problems. The solutions to these problems are then compared to identify the most promising approach.

The second question is how an LTE system can be modified to dynamically adapt to the availability of compute resources. The thesis proposes a heuristical approach which makes efficient use of the available resources, does not introduce significant complexity by itself, and is easily integrated into existing system architectures. The design of this system is based on findings from the previously solved optimization problems.

The third task is to show the efficiency of the proposed system. In this thesis, the evaluation is performed by simulation of the proposed system in a modeled LTE system. The performance achieved under different limitations of compute resources is compared to the performance achieved in an unlimited system. In addition, the proposed system is compared to different optimization problems to assess the impact of the simplifications taken during system design.

Summarizing, this thesis proposes a mechanism to make the compute resource utilization of a C-RAN system elastic. At the cost of small reduction of the network performance, the system can adapt the computational complexity to the available compute capacity. Its high efficiency is substantiated in sound simulation studies.

## 1.5 Outline of the Thesis

This thesis is structured as follows. The proposed mechanism for elastic utilization of compute resources is developed and evaluated using the example of an LTE mobile communication system. Therefore, chapter 2 introduces LTE and describes the most relevant aspects required for this thesis. It first gives an overview over the LTE architecture and explains the concept of C-RAN. Subsequently, it describes the standardized mechanisms to achieve efficient operation of the radio channel. Finally, the framework for resource management is described, which serves as a foundation for the following chapter.

This thesis evaluates the influence of link layer mechanisms, *resource allocation* (RA), and interference on the required processing effort. While the link layer mechanisms are already covered by chapter 2, the remaining aspects are targeted by the following two chapters.

Chapter 3 focuses on the allocation of radio resources. This is not covered by the LTE standardization, but allows vendors to differentiate from competitors. The chapter first discusses objectives and performance metrics from the perspectives of a single user and of the network provider. RA can be interpreted as optimization problem or tackled with a heuristic. Thus, the chapter provides an overview over the relevant literature for both approaches.

The efficient operation in multi-cell environments is targeted in chapter 4. Here, the main focus is on the handling of interference, especially doing so by coordinating multiple neighboring transmitters. This has previously been approached by defining and solving optimization problems. Thus, a structured comparison of approaches from literature is provided by mapping their definitions to a unified resource model. Finally, the chapter also gives an overview over advanced mechanisms, where a single user is simultaneously served from multiple sites.

Chapter 5 represents the central chapter of this thesis. It begins with an in-depth motivation and a detailed discussion of the research questions and contributions. Subsequently, it provides a classification of related work from three different subject areas. It then defines a common system model used for all evaluations in this thesis. To gain insight into the system and understand the interrelationships of relevant variables, different variants of an optimization problem are defined and evaluated. The proposed mechanism for elastic utilization of compute resources is then designed based on the findings of these evaluations. The chapter concludes with a discussion of the properties of this mechanism.

The main requirement for the proposed mechanism is to maintain high network performance in situations where only limited compute resources are available. This is evaluated in detail in chapter 6. First, the performance of the proposed heuristic is compared to that achieved by solving differently constrained optimization problems. As the optimization requires a simplified system model, this first evaluation is complemented by a second one with a more dynamic model. There, the performance of the proposed system is evaluated by simulation in conjunction with a realistic data traffic model.

Finally, chapter 7 summarizes the thesis and draws conclusions.

# 2 LTE Network Architecture and Mechanisms

This thesis investigates the elastic utilization of compute resources for mobile communication, using LTE as exemplary application. Although the presented mechanism is in principle applicable to other communication standards, it is designed to be used in an LTE eNodeB. Thus, some design decisions depend on conditions given by the LTE standardization.

This chapter serves as an overview of the LTE network. First, section 2.1 gives a general introduction to LTE. Subsequently, the LTE network architecture is presented in section 2.2. The focus of this thesis is on the processing performed in the BS that is related to communication with the associated mobile terminals. Therefore, the protocol stack used between these entities and the realization options for BSs are emphasized in that section.

The processing effort investigated in this thesis is caused by signal processing operations. Thus, section 2.3 provides an overview over the mechanisms that are used in LTE to achieve a high and reliable throughput on each single radio link.

While the signal processing itself can be optimized separately for each link, the BS is also responsible for the allocation of channel bandwidth to competing users. Thereto, the LTE standard comprises an elaborate framework for management of channel resources. This framework is described in section 2.4. This section does, however, not include the mechanisms that decide which resources are allocated to which user. These mechanisms are not covered by the LTE standard. They are therefore described separately in chapter 3.

## 2.1 Introduction to LTE

LTE, abbreviation for *Long Term Evolution*, is a standard for cellular communication networks. LTE networks are used commercially to deliver voice and data service to customers. Strictly, LTE describes the radio access part of the network, while *evolved packet core* (EPC) (also System Architecture Evolution, SAE) comprises the core network. Together, LTE and EPC are known as *evolved packet system* (EPS). LTE has been developed by the 3GPP, a consortium of network operators and vendors, as an evolution of previous mobile communication standards. Later versions of the standard are called LTE-A (from Release 10) or LTE-A Pro (from Release 13).

### 2.1.1 Cellular Networks

In a cellular network, a large area is split up geographically into cells. Mobile terminals inside a cell are served by the same BS over a wireless channel. In general, neighboring BSs operate independently of each other. The cellular concept was first documented by Ring [Rin47]. A *backhaul* network connects the BSs to the fixed network and also allows for communication among BSs. One application for such communication is the handover procedure which is performed when a terminal moves from one cell to another. Modern networks do also facilitate that a single user is served jointly by adjacent BSs. In the context of LTE, this is termed *coordinated multi-point* (CoMP).

The evolution of cellular networks took place in generations, each being a paradigm shift to support new demands. In the first generations many competing standards existed. As these were often applied on a national level, international roaming was difficult or often impossible. Later generations focused on global mobility, leading to common standards and harmonized frequency allocations. The following paragraphs give an overview of the evolution of cellular networks.

The first generation (1G) of commercial cellular networks started operation in 1979 with a service offered by *Nippon Telegraph and Telephone* (NTT).[1] These networks used analog modulation and provided only circuit switched voice communication.

The second generation (2G) switched to digital communication, mainly to increase the capacity for voice calls. Deployment of 2G networks started in 1992 with GSM. During the lifetime of the 2G networks, the demand for data communication increased. Consequently, the standards were extended to support higher data rates and packet switched communication.

The cellular networks of the third generation (3G) comply with common specifications by the *International Telecommunications Union* (ITU), called IMT-2000.[2] Besides other standards, this generation encompasses UMTS (*Universal Mobile Telecommunications System*) by 3GPP and cdma2000 by 3GPP2.[3] These networks have been developed to provide multimedia and high speed data services, but still support circuit switched voice services. Later extensions focused on higher data rates and reduced latency for packet communication (e. g. *High Speed Packet Access* (HSPA), cdma2000EV-DO). For an overview of IMT-2000 standards see [M.1457-12].

IMT-Advanced, the follow-up specification of IMT-2000, is generally considered as the definition of the fourth generation (4G). Officially, LTE-A and WirelessMAN-Advanced (IEEE 802.16m-2011 / 802.16.1-2012) have been approved to fulfill the requirements of this specification [M.2012-0]. However, other systems such as previous versions of LTE are also promoted as 4G. Main focus of the development was to support worldwide roaming and high data rates with a cost-efficient network. 4G networks do not support circuit switched operation. To supply voice services to the customers, the operators rely on legacy networks or apply *voice over IP* (VoIP).

The fifth generation of mobile networks (5G) is subject of current research and standardization. The requirements for the following generation are expected to be even higher data rates, lower

[1] Other standards belonging to this generation are *Nordic Mobile Telephone* (NMT), C-Netz, and *Advanced Mobile Phone System* (AMPS).

[2] IMT stands for *International Mobile Telecommunications*.

[3] Similar to 3GPP, 3GPP2 is a collaboration project developing standards for mobile communication.

latency, and improved reliability. In addition, a high number of devices with diverse requirements with respect to network performance, cost efficiency, and energy consumption should be supported. An overview of requirements, approaches, and research activities is given by Andrews et al. [And+14]. Standardization in 3GPP started with a workshop in September 2015 [Flo15]. The fifth generation will also be specified by ITU under the term IMT-2020 [M.2083-0].

### 2.1.2 LTE Releases and Features

LTE is specified by the 3GPP in consecutive releases. This section gives an overview of the most important aspects of the development. For each release, 3GPP provides a summary document on their website [3GPPReleases]. An overview of the releases up to Release 12 is given by [Cox14]. Regularly refreshed overviews are also provided by 5G Americas (renamed in 2016, formerly 4G Americas) [4GAm13; 4GAm15; 4GAm15b; 5GAm17].

The first release of LTE is Release 8, which was frozen in December 2008 [3GPP 21.101 v8.4.0]. The first LTE network has been made available to customers by TeliaSonera in Stockholm and Oslo on December 14, 2009 [Tel09]. Later releases have added features and increased performance, but in general are backward-compatible.

In comparison to earlier standards, LTE simplified the network architecture. An LTE eNodeB handles the data plane of the connected users autonomously. LTE uses multi carrier modulation. This allows to use large bandwidths of up to 20 MHz without requiring complex equalizers in the receivers. To increase data rates and reliability, LTE applies MIMO techniques with up to four antennas on both sides of the radio transmission. Power consumption in the mobile terminals is reduced by using simpler *uplink* (UL) transmissions and by application of different sleep states.

Release 9, frozen in December 2009, added location services and another mode for DL MIMO operation. In addition, it extended specifications for *Multimedia Broadcast Multicast Service* (MBMS) and home eNodeB, which were omitted from the previous release.

Release 10, the first termed LTE-A, was frozen in March 2011 and introduced a bunch of extensions to let the system conform to IMT-Advanced requirements. It allows to aggregate multiple carriers to increase the total system bandwidth. The maximum number of antennas is doubled, so that BSs and terminals can be equipped with up to eight antennas. Restrictions for UL transmissions, initially introduced to limit the complexity of mobile terminals, are partially loosened: Release 10 allows simultaneous UL transmission on non-contiguous frequencies and with multiple antennas. Cell edge performance is improved by introduction of relays. To support heterogeneous cell layouts, new mechanisms to coordinate and reduce interference are introduced.

In Release 11, various CoMP modes allow a terminal to be served by multiple base stations simultaneously. Introduction of new control resources and other extensions make the system more flexible and better suitable for many devices with low data rates, e. g. as used for machine-to-machine communication. Mobile terminals can now give hints when they expect no data transmission and can be sent to a sleep state.

Release 12 enhances MIMO modes and the associated measurement and reporting mechanisms. The changes target more efficient operation of small cells. This includes dual connectivity, where

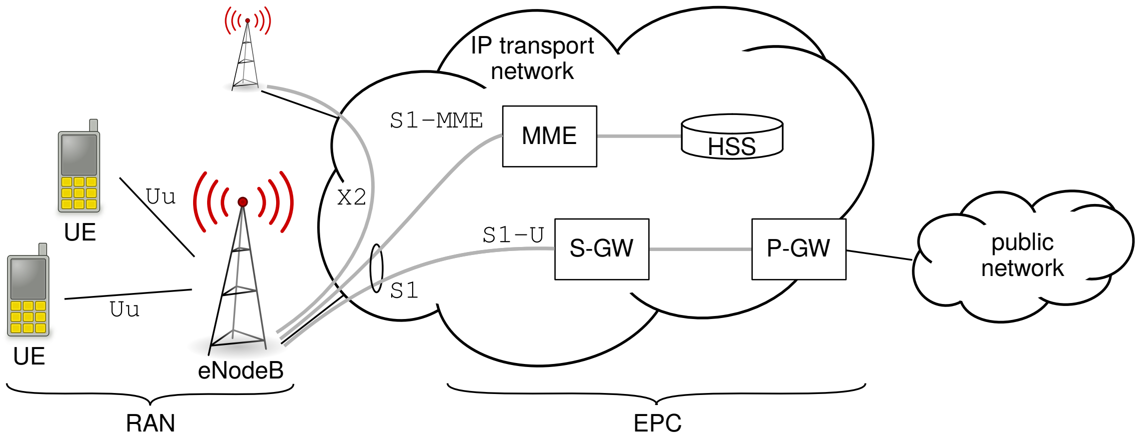


**Figure 2.1:** LTE network architecture

a mobile terminal can be connected to two cells at the same time. In addition, the release contains extensions for the connectivity of low cost and low power nodes. Proximity services allow direct communication between mobile devices. This can be coordinated by the network, but does also work if the network is not available.

In Release 13, frozen in December 2015, the official name of the system was updated to LTE-A Pro. Several modifications extend LTE for higher number of antennas for MIMO operation. In addition, multiple extensions target machine-to-machine communication and mission-crictical voice transmission for public safety operations. The system can now also operate over unlicensed spectrum, i. e. spectrum typically used for WiFi (IEEE[4] 802.11).

Release 14 was frozen in June 2017. In this release, the work focused on further enhancing technologies introduced in previous releases. The topics include the aggregation of carriers from differend bands, more antennas for MIMO, and operation over unlicensed spectrum. Furthermore, the communication between mobile devices now supports requirements of road vehicles. Public safety features are extended to allow mission-critical video and data transmissions. In addition, studies were conducted to prepare for the next generation of mobile networks.

Work on Release 15 started in the second half of 2017. This release will introduce a new radio access technology to complement the existing LTE radio access. In addition, the flexibility of the system will be increased by network slicing. This technique allows slices of the network to be optimized for use cases with contradicting requirements. Release 15 resembles the transition to the fifth generation of cellular networks.

## 2.2 LTE Architecture

LTE consists of the EPC and the RAN (also termed evolved UMTS terrestrial radio access network, E-UTRAN). The RAN comprises the base stations (termed eNodeBs) and mobile terminals (termed user equipment, UE). The EPC comprises the core network and all nodes responsible for signaling and for connecting the mobile terminals to external networks. The overall architecture is depicted in figure 2.1.

[4] IEEE stands for *Institute of Electrical and Electronics Engineers*.

The following sections highlight different aspects of the LTE architecture. First, section 2.2.1 gives an overview over the whole EPS, including the EPC and the RAN. Subsequently, section 2.2.2 focuses on the protocol stack of the radio interface. Section 2.2.3 describes different types of cells. Finally, section 2.2.4 introduces the concepts of centralization and virtualization.

### 2.2.1 Architecture of the Evolved Packet System

To simplify the system, the EPS is based on an all-IP transport network. This includes the connections between RAN and EPC and connections inside the EPC. Therefore, all connections mentioned in the further text are virtual (i. e. logical relationships), not dedicated links.

Neighboring eNodeBs are interconnected via `X2`. This interface is used for handover operation. In addition, eNodeBs exchange information to balance their load and to optimize interference management. The `S1` interface (also `S1-U` and `S1-MME`) connects eNodeBs with the EPC.

The EPC consists of nodes which handle the users' data (data plane) and nodes managing authentication, mobility and related tasks (control plane). The *Packet Data Network Gateway* (P-GW, also PDN-GW) and the *Serving Gateway* (S-GW) form the data plane of the network. The P-GW is the router that connects the EPC to an external network such as the Internet. The S-GW is an intermediate router. It routes the traffic for a UE to that eNodeB where the respective UE is connected to. The IP addresses assigned to the UEs cannot serve as locator in the network, because the IP address is constant while the UE moves. Therefore, the EPS uses tunneling to route the data packets.

The *Mobility Management Entity* (MME) handles most of the control plane communication (e. g. authentication of the UEs, tracking of UEs in the network, managing tunnels for data traffic). Multiple MMEs can be responsible for different areas. The *Home Subscriber Server* (HSS) mainly consists of a database holding subscription-related information.

The EPC encapsulates data from or to a UE in one or multiple *bearers*. Data traffic is mapped to bearers by packet filters in the UE and in the P-GW. A bearer comprises of tunnels in the EPC and the *radio bearer*. A set of *quality of service* (QoS) requirements is associated to each bearer.

The LTE RAN is specified by [3GPP 36.300] and documents referenced therein. UEs and eNodeBs are connected by the interface `Uu`, which is also termed *air interface*. A cell is defined as a set of physical resources for data transmission. UEs recognize a cell via an identifier that the network broadcasts over a geographical area. An eNodeB can serve multiple cells (e. g. sectors).

Release 12 introduces direct communication between UEs (*Sidelink*). This is mainly intended for public safety services and not further covered by this thesis. LTE supports relaying to extend coverage of the network and improve service quality. A relay node behaves like an eNodeB to terminals connected to it. Each relay node is connected to an eNodeB via a modified version of the `Uu` interface, called `Un`. Relaying is not discussed here.

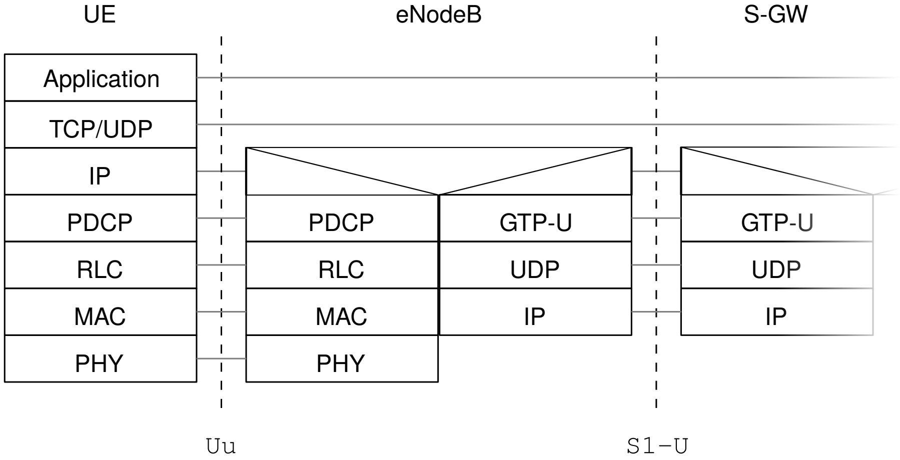


**Figure 2.2:** User plane protocol stack in an eNodeB [3GPP 23.401]

### 2.2.2 Protocol Stack of the Radio Interface

The eNodeB interconnects between UEs and Network (interfaces `S1` and `Uu`). It also performs RA and controls the parameters of the communication. The protocol stack for the user plane is drawn in figure 2.2. As the focus here lies on the eNodeB, P-GW and other nodes are omitted.

The protocols are here described for DL operation. In general, operation in UL direction is similar. The LTE *PHY* layer (layer one in the OSI model, where OSI stands for *Open Systems Interconnection*) is responsible for *forward error correction* (FEC), modulation, multi antenna operation, and mapping to physical resources. An overview of the specification of the PHY layer, including references to the detailed specification, is provided by [3GPP 36.201].

The second OSI layer consists of three sublayers: *Medium Access Control* (MAC), *Radio Link Control* (RLC), and *Packet Data Convergence Protocol* (PDCP). The MAC sublayer multiplexes MAC *service data units* (SDUs), performs fast and efficient retransmission, allocates resources, and selects parameters for the PHY layer. It is specified by [3GPP 36.321]. The RLC sublayer performs segmentation and concatenation. In addition, it is responsible for reliable in-sequence delivery of the data, which includes elimination of duplicates and a second layer of more reliable retransmission. It is specified in [3GPP 36.322]. Finally, the PDCP sublayer performs header compression and ciphering on the level of IP packets. It also handles retransmissions and duplicate detection for handover. The PDCP sublayer is specified in [3GPP 36.323]. A detailed description of the protocols is given by Dahlman, Parkvall, and Sköld [DPS14].

The control plane uses the same stack of the `Uu` interface up to and including PDCP. On top of that, the *Radio Resource Control* (RRC) protocol operates between UE and eNodeB [3GPP 36.331]. It broadcasts system information, establishes and releases RRC connections, handles paging, and controls handover. It also forwards control messages between UE and MME [3GPP 24.301].

The service access points between RLC, MAC, and PHY layer are defined as channels. Logical channels interface between RLC and MAC. They describe the type of information transferred. Transport channels interface between PHY and MAC. They specify how information is transferred over the radio interface. Physical channels directly relate to time-bandwidth resources. Tables 2.1, 2.2, and 2.3 list logical, transport, and physical channels, respectively, and describe their usage. The mapping of different channels to each other is depicted in figures 2.3 and 2.4.

Most of the channels defined by LTE are dedicated to different types of control messages. DL unicast traffic is carried by the *Dedicated Traffic Channel* (DTCH), the *Downlink Shared Channel* (DL-SCH), and the *Physical Downlink Shared Channel* (PDSCH). For UL traffic, *Uplink Shared Channel* (UL-SCH) and *Physical Uplink Shared Channel* (PUSCH) replace their respective counterparts. These channels use the major part of the time-bandwidth resources and cause the major compute effort. The further discussion therefore focuses on them.

### 2.2.3 Classes of Base Stations and Cell Layouts

The typical LTE base station is a *macro cell*. From a cell tower higher than the average rooftop height, it serves a radius of 500 m to 1000 m. It can be used to provide coverage for a certain area with minimum number of cells. Multiple *sectors* (typically three) are often used to increase capacity. Sectors share the same tower, which is in this context also termed *site*, but are differentiated by directional antennas. Sectors operate as independent cells, which means that they have independent control channels.

An LTE network has to cope with inhomogeneous demand density. Hot spots are caused either by high user density (e. g. in pedestrian areas, shopping malls, populated streets), or by groups of users making more than average use of mobile services (e. g. while waiting at bus stop or train station). To serve this inhomogeneous demand, a network operator can increase the density of the macro sites or add more sectors. However, it is often more efficient to deploy smaller cells in those areas (called *micro cell* or *pico cell*). They are in general equivalent to macro cells, except that they transmit with lower power and that antennas are often mounted lower than the average rooftop height. The combination of macro cells and smaller cells is called *heterogeneous network*. A heterogeneous network has some special requirements regarding interference and load balancing. Thus, LTE standardizes some mechanisms dedicated for these networks.

A *home eNodeB* or *femto cell* describes an even smaller cell. It is mounted by an end customer or a third party (e. g. in office buildings or shopping malls), so the network operator cannot plan its deployment. Its use can be restricted to selected UEs. A home eNodeB is connected to the EPC via non-dedicated links with varying performance (e. g. digital subscriber line, DSL). Therefore, the operation of home eNodeBs is limited w. r. t. some aspects (e. g. handover). Home eNodeBs are not further regarded in this thesis, because they are not connected to a C-RAN.

### 2.2.4 Cloud Radio Access Network

C-RAN was proposed as alternative to the classical distributed deployment of self-contained eNodeBs by IBM and China Mobile [Lin+10; CMRI11]. The abbreviation C-RAN can stand for *centralized RAN* or *cloud RAN*. Cloud RAN can be seen as an extension of the centralized RAN. Both concepts do not modify the interfaces specified by 3GPP.

The basic architecture of a centralized RAN is depicted in figure 2.5. The functions of an eNodeB are split into two groups. The lower layers are handled by a unit termed RRH. It contains ADCs, DACs, and power amplifiers and connects to the antennas. The processing of the higher layers is performed by a BBU. That is responsible for all signal processing, encoding and decoding, and higher layer protocols. Multiple BBUs are placed in a central office. RRHs and BBUs are

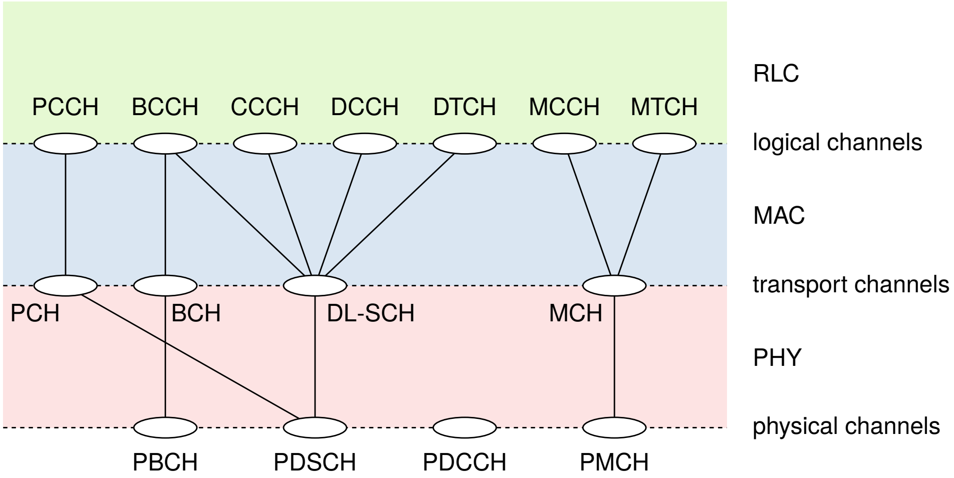


**Figure 2.3:** Mapping of DL logical channels to transport channels and physical channels (omitted for simplicity: EPDCCH, R-PDCCH, PHICH, PCFICH) [DPS14]

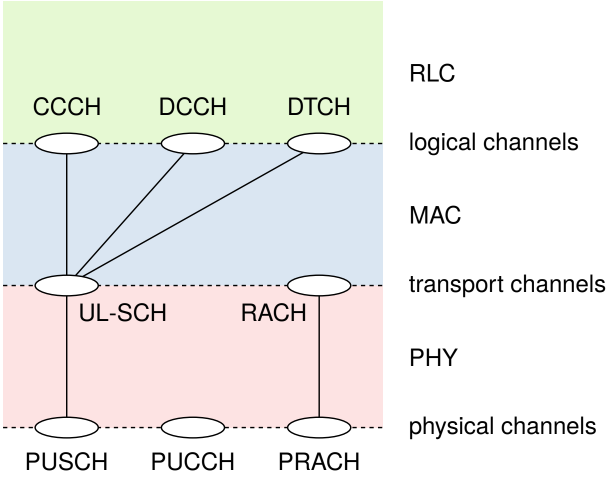


**Figure 2.4:** Mapping of UL logical channels to transport channels and physical channels [DPS14]

**Table 2.1:** Logical channels defined in LTE

| logical channel name | | description | DL | UL |
|---|---|---|---|---|
| *Broadcast Control Channel* | BCCH | broadcast of system control information | x | |
| *Paging Control Channel* | PCCH | paging information, notifications in case system information changes | x | |
| *Common Control Channel* | CCCH | control channel used for UEs without RRC connection | x | x |
| *Dedicated Control Channel* | DCCH | control channel used for UEs with RRC connection | x | x |
| *Multicast Control Channel* | MCCH | control channel for MBMS related information | x | |
| *Dedicated Traffic Channel* | DTCH | for data plane traffic, unicast | x | x |
| *Multicast Traffic Channel* | MTCH | for data plane traffic, multicast | x | x |

**Table 2.2:** Transport channels defined in LTE

| transport channel name | | description | DL | UL |
|---|---|---|---|---|
| *Broadcast Channel* | BCH | fixed, predefined modulation and coding | x | |
| *Downlink Shared Channel* | DL-SCH | flexible RA, modulation, and coding; also used for broadcasting | x | |
| *Paging Channel* | PCH | for paging | x | |
| *Multicast Channel* | MCH | multicast | x | |
| *Uplink Shared Channel* | UL-SCH | flexible RA and encoding | | x |
| *Random Access Channel* | RACH | special encoding for limited information, collisions possible | | x |

**Table 2.3:** Physical channels defined in LTE

| physical channel name | | description | DL | UL |
|---|---|---|---|---|
| *Physical Broadcast Channel* | PBCH | used to broadcast information required by UEs before they access the network | x | |
| *Physical Control Format Indicator Channel* | PCFICH | carries information required to decode the PDCCH | x | |
| *Physical Downlink Control Channel* | PDCCH | carries information required to decode the PDSCH and transmit on the PUSCH (RA, PHY layer parameters) | x | |
| *Enhanced PDCCH* | EPDCCH | allows for more flexible encoding than PDCCH; introduced in Release 11 | x | |
| *Relay PDCCH* | R-PDCCH | PDCCH dedicated to relaying | x | |
| *Physical Downlink Shared Channel* | PDSCH | carries unicast data, including user data and higher layer control information | x | |
| *Physical HARQ Indicator Channel* | PHICH | used to give UEs feedback regarding the decodability of UL transmissions | x | |
| *Physical Multicast Channel* | PMCH | carries multicast data | x | |
| *Physical Random Access Channel* | PRACH | used for the random access procedure by non-synchronized terminals | | x |
| *Physical Uplink Control Channel* | PUCCH | used to transmit control information in UL, including measurement reports, requests for scheduler grants | | x |
| *Physical Uplink Shared Channel* | PUSCH | carries all other data in UL direction | | x |

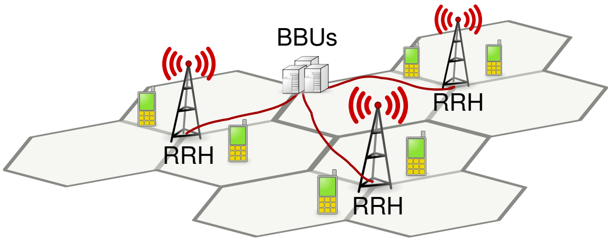


**Figure 2.5:** C-RAN architecture

connected via high speed links. Due to the high requirements w. r. t. throughput and latency, the most prominent approach is to realize these using the *Common Public Radio Interface* (CPRI) on top of a direct optical link [Oli+16].

The exact functional split between RRH and BBU is not fixed. The EU project *iJOIN* evaluated different options for this split [Wüb+14; Ros+14; Wüb+15]. Performing the split directly between PHY layer and ADC/DAC allows to centralize all signal processing, but results in high bandwidth requirements for the interconnection. Splitting higher in the protocol stack can allow to cope with lower bandwidth for the interconnection. However, it requires that the RRHs are equipped with signal processing capabilities. It is here assumed that the BBUs perform at least the MAC layer and user-specific PHY layer operations (e. g. modulation, coding, and MIMO processing).

The centralization of the BBUs already provides multiple benefits. They can share infrastructure such as housing, power supply, cooling facilities, and backhaul link. In a central office, the BBUs are easier to access for maintenance. Furthermore, CoMP becomes easier to realize, because high performance interconnection between BBUs can be realized easily. The obvious drawback of this architecture is the requirement of optical links between RRHs and BBUs. It is therefore most suitable for newly installed sites. However, it can also be reasonable to deploy C-RAN for older sites, when those have slow backhaul links which have to be upgraded anyway.

Cloud RAN extends the concept of centralized RAN by adopting principles from IT clouds. Thereto, the co-located BBUs are realized as virtual instances that run on a shared pool of compute hardware. By sharing compute hardware, the accruing load is multiplexed. This leads, first, to smoothing of short-term fluctuations, and, second, to balancing between differently frequented areas. Both effects allow to operate compute resources at a higher utilization. It is therefore expected that the shared pool requires less processing power than the sum of that required by the equivalent self-contained BBUs [I+14a; WGP13].

C-RAN is often discussed in conjunction with a software implementation of the BBUs running on general purpose hardware [I+14b; Wüb+14]. This is, in principle, equivalent to *software defined radio* (SDR), although that term is used more often to describe prototyping platforms [Skl+16]. Note that, although that is not in focus here, the combination of virtualization with specialized hardware is also possible.

Traditional eNodeB implementations rely on *application-specific integrated circuits* (ASICs), *field-programmable gate arrays* (FPGAs), and *digital signal processors* (DSPs) to perform the numerically complex calculations. Switching those to *general purpose processors* (GPPs) can bring multiple benefits. General purpose hardware can be bought as complete system. It thus does not require any effort for hardware engineering. This saves cost and allows to develop products more quickly. In addition, generic hardware is more flexible than specialized hardware.

Modifications of the BBU software allow to change how the underlying hardware is used after that is deployed. A comparison of DSPs and GPPs for BBU operation is performed by Checko et al. [Che+14]. That publication also provides an overview over prototypical GPP implementations.

When BBUs are executed as software programs on general purpose hardware, it seems obvious to also use common *operating systems* (OSs) and virtualization software. However, the LTE air interface imposes RT requirements on its implementation. These are commonly seen as a challenge for the application of IT techniques to C-RAN [I+14a; Ros+15a; Zho+16; PHT16].

A subject closely related to C-RAN is *network function virtualization* (NFV). It describes a similar approach, but is mostly focused on the functions of fixed networks. Another related subject is *software defined networking* (SDN). It does not stand for an all-software implementation of network functions, but is rather concerned with the combination of high-throughput specialized hardware with a flexible control function implemented in software.

Summarizing, C-RAN can be seen as the combination of centralization, software implementation, and virtualization. C-RAN is seen as an important building block for future networks such as 5G [Ros+15a; 5GPPP16]. An all-encompassing discussion of approaches, challenges, and benefits, together with further references, is provided by Checko et al. [Che+14].

## 2.3 Efficient Operation of the Radio Channel

The amount of spectrum available for cellular communication is limited. Licenses to use spectrum are expensive. Thus, the efficient utilization of the available spectrum is a major objective for the design of commercial radio networks. In LTE, multiple mechanisms cooperate to facilitate this. First, modulation and coding are used to make best use of a single radio channel. Second, MIMO techniques extend this by simultaneously using multiple spatially separable channels. Third, opportunistic RA exploits differences in the fluctuations of the radio channels to multiple UEs. Even with these mechanisms, the available spectrum is still considered a scarce resource. Thus, the resources have to be managed carefully to balance the interests of all participants.

This section targets the efficient operation of the radio channel. It focuses on the techniques applicable to a single communication link. The following chapter 3 is concerned with objectives and mechanisms for RA. Subsequently, chapter 4 covers mutual interference and coordination between cells.

This section is structured as follows. Initially, section 2.3.1 gives an overview over the physical effects impacting the radio channel. Subsequently, sections 2.3.2 and 2.3.3 target modulation and error correction, respectively. Section 2.3.4 introduces techniques for multi-antenna operation. Thereafter, section 2.3.5 gives a short introduction to opportunistic RA. As all these techniques require information about the radio channel, section 2.3.6 covers the mechanisms to measure and communicate its characteristics. Section 2.3.7 summarizes all techniques targeting a single communication link and gives an overall picture of the necessary processing in an eNodeB.

### 2.3.1 Characteristics of the Wireless Channel

The wireless channel encompasses physical effects between transmitter and receiver, which often lead to non-perfect reception of the transmitted signal. The transmitted signal is affected by attenuation, reflection, and diffraction. The channel effects are typically classified according to the spatial correlation of the variations. This spatial correlation maps to a temporal correlation when a UE moves. A detailed overview of the physical effects and typical models is given in [Stü11]. Here, the channel effects are described for the DL direction. The description also applies analogously for the UL direction.

First, the received signal power depends on the distance between BS and UE. The farther the UE moves from the BS, the stronger is the attenuation and the lower is the received signal power. This effect is called pathloss. Multiple models for the pathloss have been created based on measurements. In general, it is described as formula with distance, carrier frequency, antenna heights, and scenario (e.g. urban, suburban or rural) as input parameters.

The UE may have a *line of sight* (LoS) connection to the BS. In that case, the signal traveling the direct path is the main component of the received signal. However, often the UE is in a *non line of sight* (NLoS) condition. In that case, a communication may still be possible because the signal from a BS is reflected and diffracted at obstacles (scatterers) in the surrounding. This leads to the signal arriving at the UE's antenna via multiple paths. When the UE moves, paths can be obstructed and new paths can become usable. This leads to variations of the received signal strength on a spatial scale of multiple meters. This effect is called shadow fading. Measurements have shown that the received signal strength at a constant average distance from the BS can be modeled as log-normal distribution.

The paths have different lengths and their signals undergo different physical effects. That leads to a different received power, phase shift, and polarization at the receiver. The relation of the phase shifts depends on frequency of the signal and position of the UE's antenna. In NLoS conditions the paths transfer similar power. Therefore, signals with different phase shifts may add up constructively or destructively. In addition, the phase of the received sum signal varies. In LTE, wavelengths are in the order of magnitude of a few centimeters. Hence, small movements of the antennas or the scatterers change the relative phases of the signals. This effect is called fast fading, as the power of the received signal varies within milliseconds. The term *coherence time* describes the duration over which the fast fading is correlated, which depends on the speed of sender and receiver.

As LTE uses a wide bandwidth of up to 20 MHz per carrier, the relative phases vary over the bandwidth. This causes the attenuation and the phase of the received signal to be frequency-dependent. The degree of variations over frequency is described by the *coherence bandwidth*. It mainly depends on the *delay spread*, i. e. the relative delay of the signals received over different paths. In case of a LoS scenario the power of the received signal is typically modeled as Rician distributed, while in an NLoS scenario it is modeled as Rayleigh distributed.

Multiple antennas of the same UE experience different, but not independent fast fading. The correlation of the fading processes depends on the distance of the antennas, their polarization, and their directionality. At the same time, the pathloss and the slow fading of the signal received at the antennas is similar. For more details on MIMO channels refer to section 2.3.4.1.

The capacity of a channel is not defined by received power alone, but by the difference between received power and the power of disturbing signals like noise and interference. Noise is a general term describing unwanted signals which are overlaid with the desired signal at the receiver. Typically, noise consists of thermal noise, atmospheric noise, noise generated in the receiver circuits, and other sources of disturbances. The noise limits the system capacity in cases where the power of the received signal is weak, e. g. in rural areas where cells are large, in coverage holes, and inside buildings. The term *signal to noise ratio* (SNR) describes the relation of received signal power to noise power.

In urban scenarios the interference from neighboring cells often has stronger impact on the channel than the noise. Adjacent LTE cells typically use the same frequency. At the border between two cells, the signals of both cells are received with the same average power. To receive data, the signals of both cells have to be distinguished, e. g. using coding or by proper correlation of signals received by multiple antennas. To include the effect from interference, SNR can be enhanced to *signal to interference and noise ratio* (SINR). Here, the signal power is divided by the sum of interference and noise powers.

The capacity of the radio channel depends on the available bandwidth and the transmit power. According to Shannon [Sha49], for an *additive white Gaussian noise* (AWGN) channel it is limited by

$$C = W \log_2 \left( 1 + \frac{S}{N} \right) . \tag{2.1}$$

Here, $C$ denotes the channel capacity in bit/s, $W$ the bandwidth in Hz, $S$ the transmit power and $N$ the noise power.

The equation shows that it is easy to increase the channel capacity by using more bandwidth. However, this is mainly a question of acquiring spectrum licenses, and not a technical problem. To measure the efficiency with which a system uses a given amount of bandwidth, the metric spectral efficiency (also *bandwidth efficiency*) is used. This denotes the channel capacity divided by the bandwidth. The objective of modulation and encoding is to maximize the spectral efficiency and thereby get close to Shannon's theoretical bound.

### 2.3.2 Modulation and Demodulation

According to Stüber [Stü11], "modulation is the process whereby message information is embedded into a radio frequency carrier." Modulation can be analog, where an analog signal (such as an audio signal) is modulated onto a carrier. In contrast, modern radio systems apply digital modulation. There, one or multiple bits of information are mapped to a symbol (i. e. an analog pulse) to be transmitted. At the receiver side, the symbols are detected and converted back to bits. As the detection can be erroneous (e. g. in case noise disturbs the signal), the transmission is typically protected by coding (see section 2.3.3).

The remainder of this section is structured as follows. First, section 2.3.2.1 introduces symbol alphabets. Subsequently, section 2.3.2.2 describes multi carrier modulation and highlights its advantages over single carrier modulation. Finally, section 2.3.2.3 focuses on the standardization of modulation in LTE.

#### *2.3.2.1 Symbol Alphabets*

A modulation scheme is based on a symbol alphabet, which maps bit sequences to analog values. Amplitude and phase of a transmitted signal can be used to encode information. These can be translated to the real and imaginary parts of a complex number.

*Quadrature amplitude modulation* (QAM) describes a symbol alphabet where real and imaginary part encode separate bits. Its symbols can be visualized as quadratic arrangement of equidistant points in the complex plane. The simplest version of QAM encodes two bits of information into four symbols. It can be termed 4-QAM, however the label *quadrature phase shift keying* (QPSK) is more common. Its four symbols are differentiated by phase only. Higher order modulation (e. g. 16-QAM with 16 different symbols) allows to transmit more bits per symbol. However, to be able to decode symbols from higher order modulation scheme, a larger SNR is required than to decode symbols from a simpler alphabet.

#### *2.3.2.2 Single Carrier and Multi Carrier Modulation*

The classical approach of modulation is that of a single carrier system. Multiple symbols are transmitted subsequently on the same carrier. The realizable symbol rate is proportional to the available bandwidth. So, to make use of a large bandwidth, the symbol rate has to be high. However, the channel causes multiple delayed versions of each symbol to arrive at the receiver. In case the duration of the symbols is not significantly longer than the delay spread of the channel, subsequent symbols interfere with each other. This effect is called *inter symbol interference* (ISI). It is possible to recover the transmitted signal using equalization at the receiver. However, if performed in the time domain, this is computationally complex for large bandwidths.

One popular approach to mitigate this is to apply multi carrier modulation (known as discrete multitone transmission, DMT or orthogonal frequency division multiplexing, OFDM) [Stü11].This approach uses digital signal processing to divide the large bandwidth into multiple orthogonal subcarriers. Each subcarrier has a low bandwidth. Therefore, symbols are transmitted using a lower symbol rate. The longer symbol duration reduces the influence of ISI. For OFDM, the inverse *discrete Fourier transform* (DFT) is calculated to transfer per-subcarrier symbols from frequency domain to time domain.[5] The result is one OFDM symbol comprising all subcarriers.

A *cyclic prefix* (CP) is prepended to each OFDM symbol. This means that the last fraction of the OFDM symbol is copied and prepended to the beginning. If the length of the CP is longer than the delay spread of the channel, the receiver can select samples where OFDM symbol overlaps only with itself. Thereby ISI is eliminated.

Different subcarriers can carry information intended for or originating from different users. The multiple-access scheme based on OFDM is called OFDMA.

To demodulate a received OFDM symbol, the receiver performs the DFT to transform the received data to the frequency domain. However, the fast fading of the channel distorts amplitude and phase of the signal. This distortion can be different for each subcarrier. To be able to revert these channel effects, the current state of the channel has to be known at the receiver. Therefore,

[5] Typically, the *fast fourier transform* (FFT) is applied, which is an efficient algorithm to calculate the DFT.

known symbols (called *pilots* or *reference symbols*, RSs) are transmitted interleaved with the data symbols. These allow the receiver to measure the channel distortions. The fact that the channel is correlated over time and frequency can be used to interpolate the channel effects for the remaining symbols [CTL12]. Reverting the channel effects (equalization) is then performed in the frequency domain. This has lower complexity than performing it in time domain.[6]

One drawback of OFDM is the high *peak-to-average power ratio* (PAPR). The superposition of independent subcarriers leads to a higher variance of the output signal than experienced in a single carrier system. This imposes special requirements for the power amplifier. To provide sufficient headroom for the rare peaks, in average the amplifier has to be operated at a fraction of its maximum power. This leads to inefficiency regarding cost and power consumption. Both factors are especially relevant in the mobile terminals.

An alternative to plain OFDM is DFT-spread OFDM. There, the transmitter creates a sequence of symbols in time domain as if it would perform single carrier modulation. A DFT then maps that sequence to the frequency domain. The output of the DFT is assigned to subcarriers and further processed like in OFDM. The same process is reverted at the receiver side. A DFT is used to transfer the received signal into frequency domain. Equalization performed there, and finally data from the relevant subcarriers is transformed back to time domain and demodulated.

The DFT-spread OFDM combines benefits of single carrier and OFDM modulation. In case the used subcarriers are adjacent, the combination of DFT and inverse DFT at the transmitter results in a frequency shift of the signal. This allows to transmit signals of multiple users in parallel on different subcarriers. The corresponding multiplexing scheme is called *single carrier frequency division multiple access* (SC-FDMA). The receiver can simply differentiate signals from different users. Also, it can perform the equalization in the frequency domain. Myung [Myu07] gives a detailed introduction to DFT-spread OFDM and SC-FDMA.

#### *2.3.2.3 Standardization in LTE*

The modulation used in LTE is specified in [3GPP 36.211]. In DL, LTE uses OFDM with a subcarrier spacing of 15 kHz. In UL, DFT-spread OFDM is applied. The standards also define an alternative subcarrier spacing of 7.5 kHz. However, that is not used for unicast data transmission and therefore not regarded in the further discussions.

The definitions of durations are based on the time unit $T_s = 1/(15000 \cdot 2048)\text{s} \approx 32.6\,\text{ns}$. The duration of an OFDM symbol without CP is $2048 \cdot T_s \approx 66.7\,\mu\text{s}$. LTE allows to configure one of two different CP durations. The normal CP has a duration of $144 \cdot T_s \approx 4.7\,\mu\text{s}$ (or 5.2 µs for some OFDM symbols). The extended CP has a duration of 16.7 µs. It allows the system to tolerate a higher delay spread. However, this comes at the expense of reduced data rates, as less OFDM symbols can be transmitted per time interval. This thesis assumes that the normal CP is used.

On the subcarriers provided by OFDM, or on the single carrier for DFT-spread OFDM, the actual modulation symbols are transmitted. The LTE data plane uses QPSK, 16-QAM, and 64-QAM to encode 2 bit, 4 bit, and 6 bit per symbol, respectively. Release 12 extended this by 256-QAM (8 bit per symbol) for DL operation. Modulation schemes used by different users are

[6] RSs are also used to decode a multi-antenna channel. This is discussed in section 2.3.4.4.

independent. To make optimal use of instantaneous channel conditions, LTE continuously adapts the used modulation scheme in DL-SCH and UL-SCH. This works together with coding and is thus explained in section 2.3.3.1.

### 2.3.3 Error Correction

Random channel effects like fading and noise make communication unreliable. A system can cope with this either proactively or reactively, or by a combination of both approaches. First, FEC can be used at the transmitter to enhance the data with redundancy information. This can compensate bit errors in a transmission. Second, incorrectly received data can be detected by calculating a checksum both at transmitter and receiver. In case the checksums do not match, the system can repeat the failed transmission.

LTE applies a combination of these approaches to achieve high reliability and throughput. On the PHY layer, FEC coding makes transmissions robust. A *cyclic redundancy check* (CRC) is used to verify correct decoding. In case decoding fails, data is retransmitted by one of two mechanisms.

The following section 2.3.3.1 describes the FEC coding mechanisms used in LTE. After that, section 2.3.3.2 gives an overview over the applied retransmission protocols.

#### *2.3.3.1 Forward Error Correction Coding*

The main idea of FEC coding is that instead of the original data, a codeword is transmitted which contains more information. The additional information provides redundancy, i. e. allows to reconstruct the original data even if part of the codeword is distorted. The fraction of raw bits divided by codeword bits is called code rate. Multiple coding techniques exist which differ with respect to performance for different channel effects and decoding complexity. Stüber [Stü11] provides an overview over such techniques.

The coding mechanisms applied in LTE are defined in [3GPP 36.212]. For user data transmitted via the DL-SCH LTE uses a turbo code with a fixed code rate of ⅓. The encoder consists of two parallel convolutional encoders and an interleaver. At the receiver, the data is decoded by an iterative algorithm. A detailed description of turbo encoding and decoding is provided by Chiueh, Tsai, and Lai [CTL12].

Data encoded with a low code rate is well protected against distortions, however the additional redundancy reduces the available capacity for original data. On the contrary, data encoded with a high code rate has low overhead, but cannot tolerate much distortion. Therefore, to achieve highest possible efficiency under different channel conditions, the code rate has to match the instantaneous channel quality. Instead of encoding with different encoders, LTE applies *puncturing* (also termed *rate matching*). There, the transmitter does not transmit the whole codeword but leaves out a part of the codeword.

In case a codeword is spread over sufficient large frequency range or time span, the involved subcarriers experience independent fast fading. By commonly encoding data that is transmitted via independently fading channel components, the system can exploit *diversity*. With sufficiently robust encoding the transmission can tolerate a low channel quality on some of the subcarriers. For decoding, it is irrelevant which subcarriers experience severe attenuation and which do not. To determine the required code rate only statistics of the channel have to be known at the receiver, not the actual realization.

The most efficient modulation scheme and code rate required for a reliable transmission depend on the channel quality. To achieve efficient transmission for different channel qualities, LTE repeatedly adapts the *modulation and coding scheme* (MCS). This is known as *adaptive modulation and coding* (AMC). It is a subset of the more general *link adaptation* (LA) mechanism. The system tries to achieve a configured target decode probability. To estimate the channel at the transmitter, the UEs report their current channel quality to their BS.[7] Based on the reports, the BS selects the MCSs for the transmissions. A control loop can be used to determine a user-dependent offset based on the experienced decode probability. This allows the BS to compensate for optimistic or pessimistic reports by the UEs.

#### *2.3.3.2 Retransmission Mechanisms*

Due to fast fading and random noise, the transmitter cannot exactly know the instantaneous channel quality. Even if the data is encoded robustly, the channel can render decoding impossible. Therefore, in wireless networks decode errors cannot be completely avoided.

To achieve a low error probability with FEC alone, large amounts of redundancy are required. This severely impacts the efficiency of the transmission. Therefore, in LTE the proactive FEC mechanism is complemented by the two reactive mechanisms *automatic repeat request* (ARQ) and *hybrid ARQ* (HARQ). These allow to (partially) repeat a transmission in case of errors. Thereby, a higher error probability for the FEC can be accepted. The main drawbacks of retransmission protocols are additional delay and overhead.

On the PHY layer, HARQ is used as efficient mechanism to repair broken transmissions. The sender transmits a FEC-encoded codeword together with a CRC checksum. The receiver decodes the codeword and validates the checksum. It then notifies the sender about the decodability. In case the receiver could not decode the codeword, HARQ provides additional information. To do so, it transmits bits that were removed by puncturing and / or repeats already transmitted bits.

Multiple variants of repeatedly received bits can be combined by the receiver. As noise is independent while the signal is not, this helps to increase the SNR. The reception of previously unknown bits of the codeword in fact reduces the code rate. In case the codeword is still not decodable with the additional information, HARQ can further extend the received information.

HARQ is implemented in LTE as a stop-and-wait protocol. Multiple HARQ processes operate in an interleaved pattern to hide the latencies introduced by signaling. For each transmitted codeword, the eNodeB can dictate whether it is to be combined with the previously received codeword of the same process or whether the old data is to be discarded.

HARQ is efficient, because useful information received from the first transmission can be reused. In LTE, the HARQ feedback signaling is designed for low latency, but also to cope with few

[7] For a detailed discussion of channel measurements see section 2.3.6.

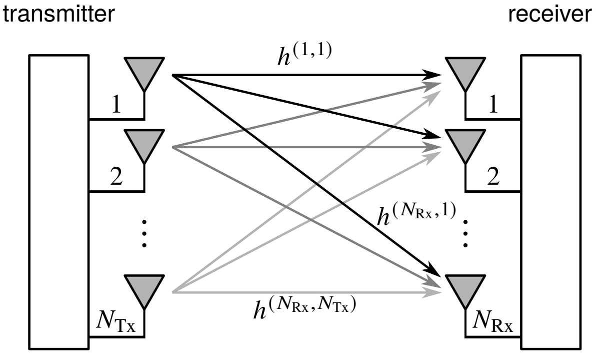


**Figure 2.6:** MIMO system with $N_{\mathrm{Tx}}$ transmit antennas and $N_{\mathrm{Rx}}$ receive antennas

channel resources. This can result in feedback errors, where the transmitter assumes that the receiver decoded the data but that was not possible. It is therefore complemented with the more reliable ARQ.

The ARQ protocol is part of the RLC functionality. RLC assigns sequence numbers to transmitted blocks of data. The receiver uses these to reorder the received blocks and to detect and eliminate duplicates. In addition, the receiver infrequently sends status reports. In contrast to HARQ feedback, these are robustly coded, and can be received reliably. By not reporting separately on every block of data, the overhead is reduced. However, this introduces additional latency.

RLC is operated in one of three different modes. In *transparent mode* RLC is not in operation. This is used for broadcast transmissions of control messages. In *unacknowledged mode* RLC performs reordering but does not retransmit data. This is used for services which prefer packet loss to additional delay (e. g. VoIP). Finally, in *acknowledged mode* ARQ is enabled. It is applied for services requiring low packet loss (e. g. those using Transmission Control Protocol, TCP).

A detailed overview of ARQ and HARQ is given by Dahlman, Parkvall, and Sköld [DPS14].

### 2.3.4 Multi-Antenna Transmission and Reception

Multiple antennas on the transmitter and the receiver add additional degrees of freedom to the system. This technology is called MIMO. When comparing to a MIMO system, a system with a single transmit and a single receive antenna is termed *single-input single-output* (SISO).

A well-structured survey of MIMO techniques is provided by Mietzner et al. [Mie+09]. They explain that MIMO has multiple benefits: First, a smaller error rate can be achieved by utilizing spatial diversity (*diversity gain*). Second, the SINR at the receiver can be improved by combining antennas and actively steering their propagation pattern (*antenna gain*). Third, the capacity of the channel can be increased by employing spatial multiplexing (*multiplexing gain*). In other literature, the antenna gain is sometimes considered as two different aspects, namely suppression of noise (*array gain*) and steering of the antenna patterns to avoid interference (*beamforming gain*) [CTL12]. In real systems, often combinations of these benefits are achieved. All of them improve the spectral efficiency of the system, either by transmitting more symbols or by allowing higher capacity modulation and coding schemes.

Although the field of MIMO is wide, this thesis focuses on those techniques used in LTE. Section 2.3.4.1, introduces the notation for the MIMO channel and discusses how transmitter and receiver gain knowledge of its realization. Following that, sections 2.3.4.2 and 2.3.4.3 give an overview over MIMO techniques applied at the transmitter and at the receiver, respectively. Finally, section 2.3.4.4 explains standardization aspects of MIMO in LTE.

#### *2.3.4.1 MIMO Channel and Channel State Information*

Assume a MIMO system with $N_{\mathrm{Tx}}$ transmit and $N_{\mathrm{Rx}}$ receive antennas as depicted in figure 2.6. The MIMO channel can be represented as complex matrix with $N_{\mathrm{Rx}} \times N_{\mathrm{Tx}}$ elements. The channel matrix $\mathbf{H} \in \mathbb{C}^{N_{\mathrm{Rx}} \times N_{\mathrm{Tx}}}$ is defined as

$$\mathbf{H} = \begin{bmatrix} h^{(1,1)} & \dots & h^{(1,N_{\mathrm{Tx}})} \\ \vdots & \ddots & \vdots \\ h^{(N_{\mathrm{Rx}},1)} & \dots & h^{(N_{\mathrm{Rx}},N_{\mathrm{Tx}})} \end{bmatrix}, \tag{2.2}$$

where each element $h^{(r,t)}$ denotes the channel from transmit antenna $t$ to receive antenna $r$.

Using this MIMO channel, the transmitter can transmit a signal vector $\mathbf{x}$ of size $N_{\mathrm{Tx}}$. The receiver receives a signal vector $\mathbf{y}$. This consists of the transmitted signal, distorted by the channel, and noise at reach receive antenna. This can be formulated as

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}. \tag{2.3}$$

The elements of the channel matrix $\mathbf{H}$ experience different fast fading.[8] Therefore channels between the antennas are not identical. Ideally, they would fade uncorrelated, but limited correlation is typically caused by the geometrical setup of the antenna arrays. The noise received at different antennas is typically assumed to be uncorrelated.

The realization of the channel is in principle unknown to the sender and the receiver. Although decoding without channel knowledge (non-coherent detection) is possible with some MIMO schemes [Mie+09], in LTE decoding relies on channel knowledge (channel state information, CSI) at the receiver. There exist MIMO mechanisms which operate without channel knowledge at the transmitter, but also others that rely on it.

The receiver can acquire CSI by evaluating reference symbols. The transmitter can get CSI either by signaling or by utilizing the reciprocity of the UL and DL channels. However, reciprocity can only be assumed if both directions share the same frequency band.[9] Measurement and reporting of the channel state introduce overhead. In addition, if the channel changes quickly, signaling latency can render CSI inaccurate or even obsolete. A detailed introduction to channel measurement and reporting in LTE is provided in section 2.3.6.

[8] Note that in some scenarios also other channel effects like pathloss and shadow fading of the elements of the channel matrix may differ. See e. g. MU-MIMO introduced in section 2.3.4.2.

[9] I. e. for TDD systems, see section 2.4.1.1.

#### *2.3.4.2 Transmitter Side MIMO Techniques*

For transmit side MIMO operations, the following discussion differentiates between those transmitting only a single data stream and those transmitting multiple data streams. Both objectives can be achieved either with taking CSI into account, or without relying on CSI.

Beamforming means that CSI is used to optimally transmit a single signal via multiple transmit antennas.[10] For each antenna, the transmitter applies a different phase shift to the signal, such that its variants sum up constructively at the receiver. Formally, beamforming can be described as multiplying the scalar signal $x$ with a complex vector $\mathbf{w}$ before transmission. With a single receive antenna, the channel matrix $\mathbf{H}$ reduces to the vector $\mathbf{h}$. This results in equation (2.3) being modified to

$$y = \mathbf{h}\mathbf{w}x + n. \tag{2.4}$$

Beamforming statistically improves the SINR in a mobile network, because the desired signals benefit from phase alignment, while interfering signals experience random phase shifts. Alternatively, instead of achieving perfect coherent combination at the receiver, the weighting vector can be designed to cause destructive combination at another receiver. This is known as *zero-forcing*.

Diversity describes the fact that under ideal conditions, the elements of the channel matrix fade independently. Without knowledge of the realization of the channel, it can be assumed that the probability that all elements simultaneously experience unfavorable fading is low. This can be used to increase the robustness of a transmission.

A multi-antenna transmitter without CSI can apply special encoding to make use of channel diversity. One such encoding has been proposed by Alamouti [Ala98]. In LTE, a similar technique is applied, which is termed *space-frequency block coding* (SFBC). Here, the first antenna transmits two symbols on adjacent subcarriers. The second antenna simultaneously transmits a different version of the same symbols. There, the symbols are swapped, complex conjugated, and the first one is negated. An extension of this mechanism (frequency-switched transmit diversity, FSTD) is applicable to four transmit antennas.

Given that the system has several transmit and several receive antennas, spatial multiplexing transmits multiple data streams in parallel. With a linear detector and favorable channel conditions, up to $l = \min(N_{\mathrm{Tx}}, N_{\mathrm{Rx}})$ *layers* or *spatial streams* can be transmitted and decoded by the receiver. This means that on the same subcarrier at the same point in time, $l$ independent symbols are transmitted in parallel. There are different possibilities to realize this.

In principle, a separate stream of data can be transmitted by each transmit antenna. The receiver then needs to decorrelate the transmitted streams. However, this is only possible if $N_{\mathrm{Tx}} \leq N_{\mathrm{Rx}}$ and the correlation of the channel matrix is sufficiently low. Thus, if $N_{\mathrm{Tx}} > N_{\mathrm{Rx}}$, some transmit antennas remain unused. Also, a significant amount of energy may be transmitted into directions other than the receivers'.

Alternatively, the transmitter can *precode* the signal before transmission. In general, the $l$ transmit signals can be mixed arbitrarily on each of the $N_{\mathrm{Tx}}$ transmit antennas. Referring to beamforming

[10] A single antenna is sufficient to receive a signal from a beamforming transmitter. However, as introduced above, the receiver can make use of additional antennas to suppress noise and interference.

as introduced above, this can also be interpreted as simultaneously transmitting multiple beams. Formally, equation (2.3) is extended by a complex weighting matrix **W** of size $N_{\mathrm{Tx}} \times l$:

$$\mathbf{y} = \mathbf{HWx} + \mathbf{n}. \tag{2.5}$$

As with beamforming, the precoding matrix **W** has to be selected such that the components of a single signal add up constructively at the receiver, i. e. their phases are aligned. In addition, interference between data streams has to be avoided. This means that the data streams have to appear orthogonal to the receiver. With ideal CSI, using the *singular value decomposition* (SVD) to decompose the channel into independent spatial subchannels results in optimal performance [CTL12]. However, ideal CSI can typically not be acquired, because fast fading changes too quickly and signaling overhead would be significant.

Precoded spatial multiplexing can be performed either with frequently signaled, quantized CSI (*closed loop precoding*), or by falling back to longer-term statistical CSI (*open loop precoding*). Closed loop precoding tries to follow the variations of the channel to approximate the performance that would be achieved by a SVD. In LTE, the receiver measures the channel and selects an entry from a predefined codebook which maximizes the capacity. This is signaled to the transmitter, together with the supported favored number of layers.

If the channel changes too quickly, open loop precoding can be used instead. Different precoding matrices are used for each subcarrier. Transmitting a single codeword with different precoding matrices achieves diversity. The probability that at least some of the applied precoding matrices are suitable for the instantaneous channel matrix is high. A strong FEC coding allows to recover the remaining symbols. For open loop precoding, the transmitter does not require information about the realization of the channel matrix. However, the receiver is required to inform the transmitter about how many layers to transmit.

MIMO can be extended to multiple users, which is called *multi-user MIMO* (MU-MIMO). For example, a BS with multiple antennas can simultaneously transmit data to multiple UEs. In principle, this can be seen as generalized case of precoded spatial multiplexing, where the channel matrix consists of the combined channels of the served users. An overview over techniques for MU-MIMO, including non-linear precoding, is given by Spencer et al. [Spe+04].

MU-MIMO can have multiple advantages. First, the BS may have significantly more antennas than the UEs. As spatial multiplexing to a single receiver is limited by the lower number of receive antennas, transmitting to multiple UEs in parallel allows to increase the number of spatial streams and thereby use the BS antennas more efficiently. Second, even if multiple antennas are available at UE side, their closed spacing often causes their received signals to be correlated. In contrast, UEs at different positions experience less correlation of their channels. Third, the BS can select a subset of UEs for parallel service. It thereby has some influence on the combined channel matrix. This can be used to maximize the orthogonality of the channels. Drawbacks of MU-MIMO are that each UE receives only part of the transmit power and that RA becomes complex, because it includes the selection of subsets of the UEs to be served simultaneously. MU-MIMO is not considered in this thesis.

#### *2.3.4.3 Receiver Side MIMO Techniques*

A MIMO receiver sees the transmitted symbols multiplied with the product of precoding matrix and channel matrix. The concatenation of precoding matrix and channel matrix can be interpreted as effective channel $\mathbf{H}_{\text{eff}}$, which distorts the transmitted symbols. The receiver has to revert this effect to decode the data.

First, assume that there is a single transmit antenna ($N_{\text{Tx}} = 1$) and multiple receive antennas. Equation (2.3) can then be reformulated as

$$\mathbf{y} = \mathbf{h}x + \mathbf{n}. \tag{2.6}$$

Here, the received signal vector $\mathbf{y}$ consists of differently distorted versions of the same transmitted signal $x$ and noise. The same is true for a beamforming transmitter. There, $\mathbf{h}$ is replaced by $\mathbf{h}_{\text{eff}} = \mathbf{hw}$, with $\mathbf{w}$ as introduced in equation (2.4).

An arbitrary linear combination of the elements of the received signal vector can be formulated as multiplication with a weighting vector $\mathbf{g}$, i. e.

$$y = \mathbf{g}\left(\mathbf{h}_{\text{eff}}x + \mathbf{n}\right) . \tag{2.7}$$

The receiver can realize diversity by always decoding the signal of that antenna which received the highest power. This means that only a single element of $\mathbf{g}$ is set to one, the others to zero. This approach is known as *antenna selection*.

Better performance is realized by coherently combining the received signals. This improves the SNR, because noise received on different antennas can be assumed to be uncorrelated. In addition, it implicitly suppresses interference, because the sensitivity of the receive antenna array is steered towards the transmitter of the desired signal. Other weightings can be used to explicitly suppress strong interferers.

Antenna selection and coherent combining are here introduced for a single transmitted signal. They can also be applied with a more complex multi-antenna transmitter, given that there are sufficient degrees of freedom left, i. e. the receiver does not necessarily need all available antennas to decode the transmitted signal.

Multiple receivers have also been developed for spatial multiplexing. The simplest one are the linear receivers called *zero-forcing* (ZF) and *minimum mean squared error* (MMSE). Here, the receiver multiplies the received symbols with a matrix, i. e.

$$\mathbf{y} = \mathbf{G}\left(\mathbf{H}_{\text{eff}}\mathbf{x} + \mathbf{n}\right) . \tag{2.8}$$

The ZF receiver inverts the channel and precoding. The matrix $\mathbf{G}$ is defined as

$$\mathbf{G}_{\text{ZF}} = \mathbf{H}_{\text{eff}}^{\dagger} = \left(\mathbf{H}_{\text{eff}}^{\mathsf{H}}\mathbf{H}_{\text{eff}}\right)^{-1}\mathbf{H}_{\text{eff}}^{\mathsf{H}}, \tag{2.9}$$

where $()^{\dagger}$ denotes the pseudo-inverse, $()^{\mathsf{H}}$ the Hermitian transpose, and $()^{-1}$ the inverse of a matrix. The ZF receiver ideally decorrelates the data streams. However, it suffers from noise and interference amplification.

The MMSE receiver also takes into account noise and interference. It thereby achieves better performance. For MMSE, the matrix $\mathbf{G}$ is defined as

$$\mathbf{G}_{\mathrm{MMSE}} = \left(\mathbf{H}_{\mathrm{eff}}^{\mathsf{H}}\mathbf{H}_{\mathrm{eff}} + \frac{\mathbf{I}}{\rho}\right)^{-1} \mathbf{H}_{\mathrm{eff}}^{\mathsf{H}}, \tag{2.10}$$

where $\rho$ is the SNR and $\mathbf{I}$ the identity matrix of appropriate size.

Even higher performance can be achieved with non-linear receivers, e. g. *successive interference cancellation* (SIC) or *maximum likelyhood* (ML). A SIC receiver iteratively decodes streams and subtracts their interference from the received sum signal. The ML receiver theoretically achieves the optimal performance by evaluating all possible combinations of transmitted symbols. However, it is computationally complex. Therefore, approximations like *sphere decoding* are sometimes implemented instead. A detailed explanation of these and additional receivers is provided by Chiueh, Tsai, and Lai [CTL12].

#### *2.3.4.4 Standardization and Implementation in LTE*

MIMO operation for DL transmissions in LTE is standardized in [3GPP 36.211] since Release 8. MIMO for UL transmissions is introduced by Release 10. The remainder of this section focuses on DL. The standard allows to use different MIMO techniques. The BS can select the one suitable for the respective channel and other conditions.

For quickly moving UEs, the coherence time of the channel is short. It is therefore difficult to acquire useful CSI. Consequently, the BS uses a transmit side diversity encoding or open loop precoding. In contrast, for slow UEs channel measurement is possible. Highest performance is then realized with beamforming or closed loop precoding. Often, in low SINR scenarios, as typical for UEs close to the cell border, it is most efficient to transmit a single data stream with high reliability, i. e. use transmit side diversity encoding or beamforming. For situations with high SINR, capacity can be increased by spatial multiplexing.

Independent of whether CSI is required at the transmitter, the receiver has to estimate the channel effects. Therefore, LTE uses reference symbols, as introduced in section 2.3.2.2. Two approaches exist in LTE to derive the effective MIMO channel.

Release 8 defines *Cell-Specific Reference Signals* (CRSs). These are transmitted over the whole bandwidth. They are independent of allocations of the channel to UEs, and thereby not affected by precoding. Each antenna transmits a predefined pseudo-random sequence of symbols. The same subcarrier is not simultaneously used by the other antennas. Thereby, the UEs can measure the channels of up to four antennas. To derive the effective channel, the applied precoding matrix has to be signaled to the receiving UE. A codebook with precoding matrices is defined to reduce signaling overhead.

In addition to CRSs, *Demodulation Reference Signals* (DM-RSs) are introduced by LTE Release 9. They are multiplexed with the data symbols before precoding. They are therefore affected by the precoding and the channel. The receiving UE can use DM-RSs to estimate the effect of the effective channel without understanding the processing performed at the transmitter. As the BS is not required to signal the applied precoding, it is not restricted to the codebooks defined in the standard, but can instead use arbitrary precoding matrices.

The actual number and configuration of antennas at the base station is not relevant to the UEs. Instead, the UEs only have to be able to understand how data is precoded. In case of DM-RSs this is possible without further restrictions. In case of CRSs, the transmission of reference symbols and the data (precoded by a precoding matrix from the codebook) has to be consistent. Therefore, LTE defines MIMO modes in terms of antenna ports, not antennas. A single antenna port can consist of multiple antennas, e. g. vertical array of separate antenna elements. The mapping from antenna ports to antennas can be seen as a second precoding.

This concept can also be used to hide antennas from legacy terminals. For example, an LTE-A BS with eight antennas can combine those to four antenna ports with two antennas each. It can use those four antenna ports to serve Release 8 UEs. Simultaneously, it can use another eight antenna ports, which now directly map to the antennas, to serve Release 10 UEs.

Antenna ports are identified by numbers. CRSs are transmitted on ports 0, 0 to 1, or 0 to 3, depending on how many antennas the BS wants to advertise to legacy terminals. The mapping from these ports to physical antennas is always constant. In contrast, ports 5 and 7 to 14 are intended to be used with DM-RSs. Their mapping to physical antennas can differ, e. g. it is adapted to the different channels of different UEs. Other ports are used for special reference symbols or multicast operation.

For efficient signaling LTE defines *transmission modes* (TMs) to be used for data transmissions on the PDSCH [3GPP 36.213]. A separate TM is semi-statically configured for each UE. While some information regarding the MIMO scheme is fixed by the TM, other configuration can be selected independently for each allocated set of resources. This allows to adapt to the instantaneous channel conditions. This adaptation, together with AMC introduced in section 2.3.3, constitutes the LA mechanism.

The TMs defined up to LTE Release 12 are listed in table 2.4. TMs 1 to 7 are contained in the Release 8 specification. TMs added in later releases are not supported by legacy UEs.

Each TM restricts the usable multi-antenna schemes. In addition, the TMs differ in terms of reference symbols used for decoding.[11] The number of transmitted layers and the actual precoding matrix (if applicable) is free to be configured separately for each allocated set of resources. All modes except TM 1 allow to fall back to SFBC. This allows reliable communication in case no up-to-date channel measurements are available. It also allows the system to recover when the situation of the UE changed so that the configured TM is no longer suitable.[12]

[11] Channel measurements are also defined differently, see section 2.3.6.

[12] An example for such recovery is when a slowly moving UE was configured in TM 4, but suddenly starts to move quickly so that accurate CSI cannot be acquired any more. In that case, the eNodeB can use SFBC transmission to transmit outstanding data and to change the TM to something more suitable, e. g. TM 3.

**Table 2.4:** Transmission modes defined in LTE [3GPP 36.213; DPS14]

| TM | description | max. num. | ref. syms. | ant. | LTE |
|---|---|---|---|---|---|
| | | layers | for demod. | ports | Rel. |
| 1 | single antenna | 1 | CRSs | 0 | 8 |
| 2 | transmit diversity (SFBC) | 1 | CRSs | 0-3 | 8 |
| 3 | open-loop codebook based precoding | 4 | CRSs | 0-3 | 8 |
| 4 | closed-loop codebook based precoding | 4 | CRSs | 0-3 | 8 |
| 5 | closed-loop codebook based precoding for MU-MIMO | 1 | CRSs | 0-3 | 8 |
| 6 | closed-loop codebook based single layer precoding (beamforming) | 1 | CRSs | 0-3 | 8 |
| 7 | non-codebook-based precoding | 1 | DM-RSs | 5 | 8 |
| 8 | non-codebook-based precoding | 2 | DM-RSs | 7-8 | 9 |
| 9 | non-codebook-based precoding | 8 | DM-RSs | 7-14 | 10 |
| 10 | non-codebook-based precoding for CoMP | 8 | DM-RSs | 7-14 | 11 |

MU-MIMO can be used to transmit data to two UEs in TM 5 since Release 8. This relies on CRSs and codebooks as defined for spatial multiplexing. Here, a separate TM is required because the UEs have to be aware of the power reduction: While each UE receives only a single layer, the transmit power is split between the two layers for the two UEs. This restriction is lifted with the introduction of DM-RSs. As the receiver is not required to understand the processing performed by the transmitter, multiple UEs can be served simultaneously without notification.

MIMO operation is part of the LTE PHY layer. It can be seen as sublayer above OFDM processing and below FEC coding and modulation. In case of spatial multiplexing, typically two codewords are modulated and spread to the layers.[13] Two codewords allow to use different modulation and coding to adapt to the possibly different capacity of the layers. Separate codewords also allow the receiver to employ SIC, i. e. decode one codeword, subtract its interference, and then decode the other codeword. In contrast, spreading a codeword to multiple layers makes use of diversity. The sizes of the codewords are chosen such that the same number of modulated symbols are transmitted on each layer. It can occur that one codeword can be received correctly while the other one is not decodable. Thus, separate HARQ processes are used for the codewords, and codewords can be retransmitted independently.

Figure 2.7 shows a schematic of the DL MIMO processing chain in a BS. It assumes a BS with eight antennas. The operations for TMs that rely on CRSs are shown on the right, those for TMs using DM-RSs on the left. Each arrow represents the flow of data for a codeword or layer. Optional codewords and layers are represented by dashed arrows.

The processing begins with modulated symbols for one or two codewords per UE. These codewords are mapped to layers (step Ⓐ), i. e. each codeword is either allocated to a layer as

[13] A single codeword is used for single-layer transmissions. A single codeword spread to multiple layers can only be used for retransmissions.

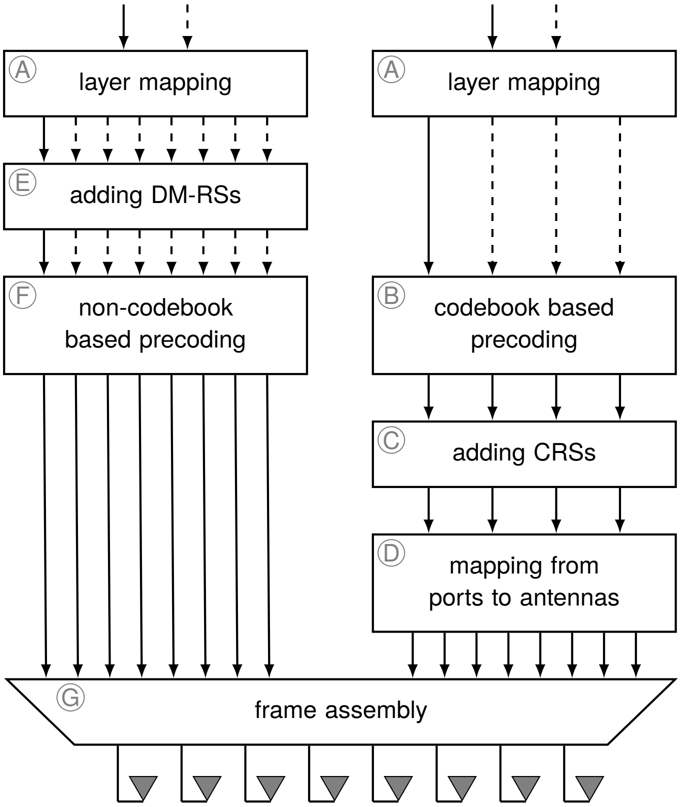

Ⓐ layer mapping
Ⓐ layer mapping
Ⓔ adding DM-RSs
Ⓕ non-codebook based precoding
Ⓑ codebook based precoding
Ⓒ adding CRSs
Ⓓ mapping from ports to antennas
Ⓖ frame assembly

**Figure 2.7:** DL transmit processing chain for MIMO operation

a whole, or its symbols are spread to multiple layers. The result are multiple layers with the same number of modulated symbols each. For the following steps, depending on the used TM, codebook based precoding or non-codebook based precoding is used.

For codebook based precoding, the symbols of the layers are now directly multiplied with the precoding matrix **W** taken from the standardized codebooks (step Ⓑ). The number of antenna ports transmitting CRSs is configured statically. The result of the codebook based precoding is always one symbol per CRS antenna port. After precoding, the precoded symbols are interleaved with CRSs (step Ⓒ). The combination of precoded data symbols and CRSs is now mapped to antennas (step Ⓓ). This can involve a second step of precoding, however in this case the precoding matrix has to be constant.

For non-codebook based precoding, the original symbols mapped to layers are first interleaved with DM-RSs (step Ⓔ). Then, data symbols and reference symbols are jointly precoded (step Ⓕ). The output of this precoding is one symbol per antenna.[14]

Finally, symbols for different UEs are multiplexed according to the desired RA (step Ⓖ). The output of the multiplexing is then forwarded to the OFDM processing, i. e. the next step will be calculation of the inverse DFT.

[14] Although not depicted in figure 2.7, CRSs are also transmitted on resources which are used to serve UEs with DM-RSs. All CRSs and all user data precoded with a matrix from a codebook are mapped to antennas by multiplication with the same constant matrix.

### 2.3.5 Opportunistic Resource Allocation

Opportunistic RA also contributes to an efficient utilization of the radio channel. The mechanisms discussed in the previous sections focus on efficient utilization of the radio channel by a single UE. In contrast, opportunistic RA targets the utilization of spectrum on cell level.

OFDM divides the available spectrum into a two-dimensional grid of subcarriers (frequency) and symbols (time). The fast fading of the radio channels of multiple UEs is uncorrelated. Thus, the capacity of the radio channels to multiple UEs differs. Furthermore, the capacity of the radio channel to one UE is different on each resource. Assume initially that these capacities are known. The BS can then assign each resource to that UE which maximizes the capacity for that specific resource. This maximizes the sum throughput of the cell. However, this approach can incur disadvantages for those UEs which have a low channel quality in general. Other approaches exist which make use of the variations but avoid discrimination.

Exact knowledge of the channel capacities cannot be acquired in reality, because they change over time. The channel parameters are measured, and the measurement serves as a prediction of the future channel quality. The accuracy of this prediction mainly depends on the velocity of the UE. Long coherence times of the channels of slowly moving UEs facilitate efficient utilization of the fluctuations. For faster UEs opportunistic scheduling cannot provide significant benefits.

Frequency diverse RA is a prominent approach to serve those UEs for which the channel quality is not known accurately. That means that resources which have a separation in frequency of more than the coherence bandwidth are assigned to each of these UEs. These resources are then combined by robust FEC coding. Thereby, the risk is reduced that all assigned resources are impacted by unexpected fading. Frequency diverse RA and opportunistic RA can be combined, so that each UE is served with the adequate mechanism.

Summarizing, opportunistic RA serves to make efficient use of the radio channels. The RA mechanism has to balance the requirements of the cell and of all served UEs. RA also has a significant influence on the compute requirements of an LTE eNodeB. The further discussion of RA is split into two parts. Section 2.4 describes the framework provided by LTE to allocate resources. Chapter 3 covers objectives and heuristics for RA, which are not standardized.

### 2.3.6 Measurement and Reporting of Channel State Information

In LTE the modulation scheme, the code rate, and the MIMO mode used for UL and DL transmissions are selected by the eNodeB. In addition, the eNodeB performs RA for both directions. Therefore, it has to be provided with CSI for both channels.

Section 2.3.6.1 gives an overview over the requirements for CSI acquisition. When a UE transmits data, the UL channel can be measured by the eNodeB itself. This is described in section 2.3.6.2. The DL channel can be measured by the UEs, and the measured information can be reported to the respective eNodeB.[15] Methods for measurement and reporting of the channel parameters are described in sections 2.3.6.3 and 2.3.6.4, respectively.

[15] For TDD operation, it is also possible to measure the UL channel only and use the reciprocity of the channel to transfer the results to the DL channel. For introduction of TDD refer to section 2.4.1.1.

#### *2.3.6.1 Requirements*

CSI is required for different components of the LTE system. AMC maximizes the capacity and limits the error rates under all channel conditions (see sections 2.3.2 and 2.3.3). It requires the eNodeB to estimate the average channel quality of each allocated set of resources. For MIMO operations, different channel information is required depending on the configured TM (see section 2.3.4). For spatial multiplexing, the eNodeB has to know the number of supported spatial layers of the channel. In addition, for closed loop precoding the precoding matrix resulting in maximum capacity has to be signaled. As this information depends on the fast fading and thereby varies over the bandwidth of a cell, it is beneficial to have this information for multiple parts of the cell bandwidth. RA algorithms can use detailed channel information to allocate those resources to UEs where the respective channel has highest quality. Finally, in multi-cell scenarios information about channels to neighboring cells can be useful to coordinate transmissions and thereby avoid interference (see section 4.2).

Ideal CSI is a common assumption for analytical and simulative evaluations, however it cannot be acquired in real systems. Instead, quantization errors and feedback delay are unavoidable sources of error. In addition, to reduce overhead for measurement and signaling, real systems limit the resolution in time and frequency domain.

Characteristics of the environment and the speed of movement of the UE determine the coherence time[16] of the channel. That defines how quickly the CSI becomes obsolete. Thereby it influences the optimal choice of measurement intervals. The optimal resolution in frequency domain depends on the coherence bandwidth of the channel.

In case the channel cannot be tracked with sufficient accuracy, the eNodeB can also work with approximate or even without CSI. A robust MCS maintains reliability with inaccurate CSI. Robust MIMO modes like SFBC work without channel information. Similar fallback approaches also exist for RA and interference avoidance. While these approaches are capable of achieving reliable transmission of data, this comes at the cost of reduced spectral efficiency. An overview over literature on networks operating with imperfect CSI is provided by Love et al. [Lov+08].

#### *2.3.6.2 Measurement of the Uplink Channel*

To measure the UL channel, the UEs transmit signals that are known by the BS. This can be either DM-RSs, which are embedded in data transmissions, or dedicated reference symbols for measurement. Using DM-RSs for UL channel measurement does not incur additional overhead. However, these transmissions do typically not cover the whole bandwidth. Therefore, LTE also allows the eNodeB to configure UEs to transmit *Sounding Reference Signals* (SRSs). Their bandwidth can be configured independently of data transmissions. They are transmitted either periodically in intervals of 2 ms to 160 ms or are triggered on demand by the eNodeB.

SRSs of multiple UEs can be multiplexed in the frequency domain. In addition, orthogonal sequences of reference symbols can be used to differentiate multiple UEs transmitting SRSs on the same subcarriers. The placement of the reference symbols in the time-frequency resource

[16] For a definition of coherence time and coherence bandwith, please refer to section 2.3.1.

grid is discussed in section 2.4.2.2. The measurement and evaluation performed by the eNodeBs is not covered by the LTE standards.

#### *2.3.6.3 Measurement of the Downlink Channel*

The measurement of the DL channel is based on CRSs or *CSI Reference Signals* (CSI-RSs). The placement of both types of reference signals in the time-frequency resource grid is discussed in section 2.4.2.1.

CRSs are transmitted on one to four antenna ports. A high density, i. e. small distances between reference symbols in time and frequency, is required because they are also used for MIMO decoding and demodulation. UEs can predict channel conditions and interference based on measurements of these CRSs. To reduce impact of noise and other sporadic effects, the measurement is averaged over time. As interference can be different in consecutive subframes (see ABSs introduced in section 4.2.4), Release 10 allows to distinguish different groups of subframes. However, CRSs come with significant overhead. Therefore, the concept is not extended to more than four antennas.

In Release 10, the newly introduced TM 9 can use either CRSs or CSI-RSs for channel measurements. CSI-RSs can be configured for one, two, four, or eight antenna ports. To reduce overhead, CSI-RSs are not transmitted every subframe, but with a periodicity of 5 ms to 80 ms. The exact position of the reference symbols can be configured flexibly.

To measure CSI-RSs, the eNodeB can configure each assigned UE differently. Each can have one set of CSI-RSs where reference symbols are received. In addition, it can ignore multiple sets of CSI-RSs. The UE then neither uses these symbols for measurement nor tries to decode data. Thus, the eNodeB can use the ignored CSI-RSs to transmit reference signals to other UEs. The symbols can also be left empty to measure the channel without interference in an adjacent cell.

TM 10, which is introduced in Release 11, always uses CSI-RSs. There, multiple independent CSI processes can be configured for each UE. Separate reports for different CSI-RSs configurations allow the eNodeB to estimate the channel under different transmission hypotheses. Also, signal strength and interference are measured on different resources. Information about *quasi-colocation* can be used to notify the UEs about which antenna ports are located at the same site and which are not (see section 4.3). This is required when measurements form multiple antenna ports are combined to estimate large-scale properties (e. g. pathloss and shadow fading).

#### *2.3.6.4 Reporting of Channel Measurements*

Reporting of channel measurements in LTE is standardized in [3GPP 36.213]. The parameters reported by a UE depend on the configured TM.

The channel quality is encoded as *channel quality indicator* (CQI). This is an index into a standardized table of modulation and coding schemes, which is reprinted in table 2.5. The UE signals to the eNodeB that combination which is expected to deliver the highest possible throughput, but can still be decoded with less than 10 % error probability.

**Table 2.5:** List of CQIs standardized by 3GPP [3GPP 36.213 v12.5.0, table 7.2.3-1])

| CQI index | modulation | code rate | bits per symbol |
|---|---|---|---|
| 0 | | *out of range* | |
| 1 | QPSK | 0.076 | 0.15 |
| 2 | QPSK | 0.117 | 0.23 |
| 3 | QPSK | 0.188 | 0.38 |
| 4 | QPSK | 0.301 | 0.60 |
| 5 | QPSK | 0.438 | 0.88 |
| 6 | QPSK | 0.588 | 1.18 |
| 7 | 16-QAM | 0.369 | 1.48 |
| 8 | 16-QAM | 0.479 | 1.91 |
| 9 | 16-QAM | 0.602 | 2.41 |
| 10 | 64-QAM | 0.455 | 2.73 |
| 11 | 64-QAM | 0.554 | 3.32 |
| 12 | 64-QAM | 0.650 | 3.90 |
| 13 | 64-QAM | 0.754 | 4.52 |
| 14 | 64-QAM | 0.853 | 5.12 |
| 15 | 64-QAM | 0.926 | 5.55 |

The optimal number of spatial layers for MIMO is termed *rank indication* (RI). For closed loop precoding, the UE selects an entry from the codebook which results in maximum channel capacity. The respective index, called *precoding matrix indicator* (PMI), is signaled to the eNodeB. Different codebooks are defined for two, four, and eight antennas. The reporting is also based on codebooks for TMs that use non-codebook based precoding.

If spatial multiplexing is used, two CQI values can be reported for the two codewords. CQI and PMI can be frequency selective, while the RI is always valid for the whole bandwidth. To reduce overhead, frequency selective reporting has a coarse granularity, which is defined as subbands.[17] Each UE reports parameters optimal for the channel conditions from its point of view. However, those can be overridden by the eNodeB.

Reporting can be either periodic or aperiodic. For periodic reporting, the eNodeB configures a reporting interval up to every 2 ms. The reports are typically transmitted on PUCCH, where the capacity for each UE is low. Therefore they are optimized for small overhead. The reports can either be wideband, i. e. valid for the whole cell bandwidth, or frequency selective. In case of frequency selective reports, the UE cycles through parts of the bandwidth. In each report, data for one of the parts is transmitted.

Aperiodic reports are requested by the eNodeB on demand, in combination with a grant to transmit the result on PUSCH. As they are not transmitted unnecessarily, and PUSCH has higher capacity than PUCCH, more detailed information can be signaled. All aperiodic reports include a wideband CQI. In addition, frequency selective parameters are reported. Either the UE selects preferred subbands and reports parameters for those, or the subbands are selected by the eNodeB.

[17] Each subband comprises of two to eight PRBs, as introduced in section 2.4.1.2.

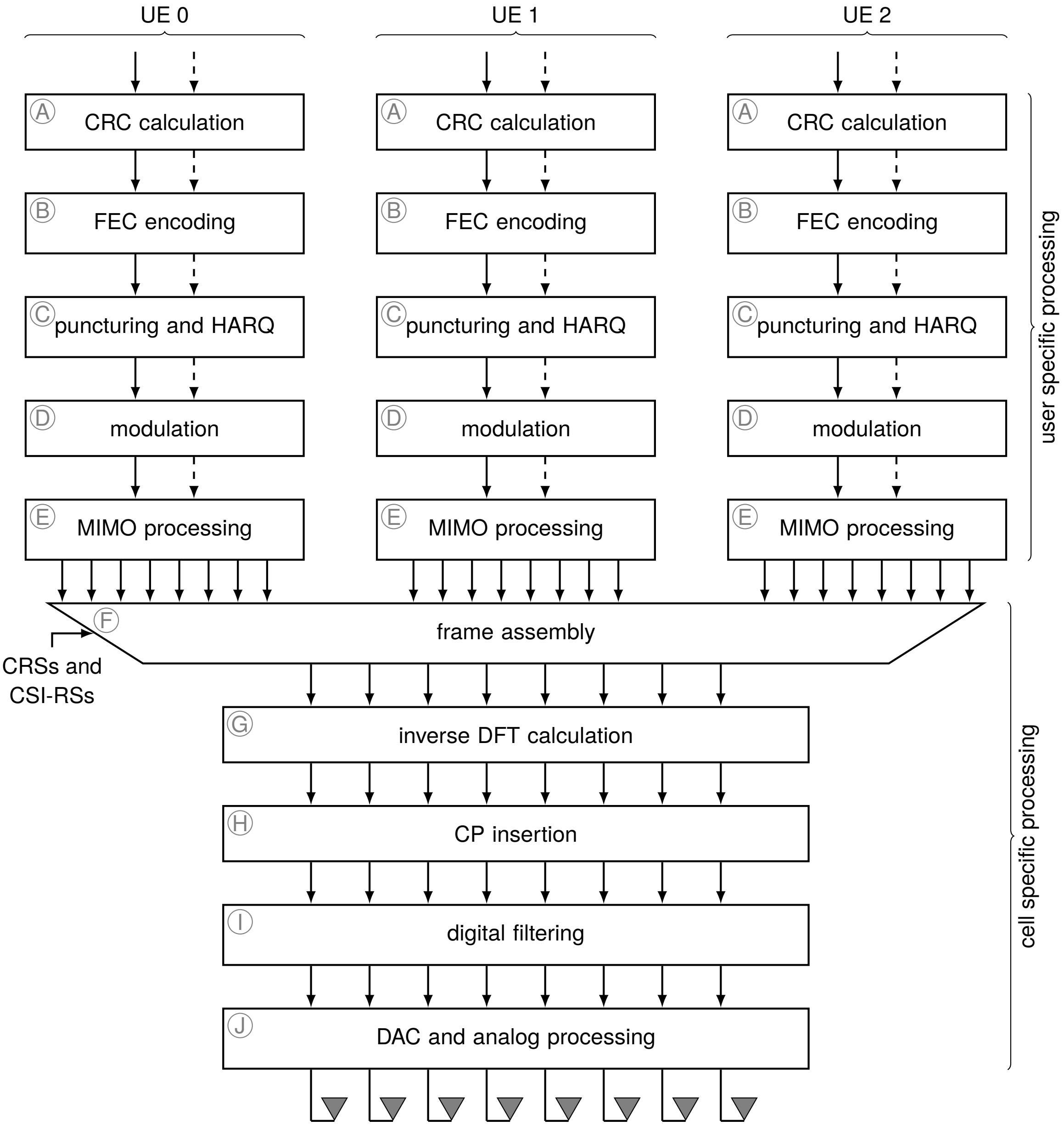


**Figure 2.8:** DL processing chain, exemplary configuration for three UEs and eight transmit antennas (partially following [DPS14])

In case the UE is configured with multiple CSI processes, the eNodeB can selectively request reports for a subset of the processes.

### 2.3.7 Summary and Overall Picture

Figure 2.8 summarizes the physical layer processing an eNodeB performs to transmit data to UEs. The first steps operate independently for each UE.

Higher layers supply the data in one or two transport blocks (depending on the applied MIMO mode). These are first protected by a CRC (step Ⓐ). This allows the receiver to detect decode errors. Data and checksums are then encoded by a turbo code with fixed rate (step Ⓑ).

Subsequently, a subset of the encoded bits is selected to match the desired code rate (step Ⓒ). The remaining information is stored to provide incremental redundancy for HARQ retransmissions. The encoded bits are then modulated (step Ⓓ) and forwarded to the MIMO processing block (step Ⓔ). That maps the modulation symbols to spatial layers, optionally adds DM-RSs, and applies precoding.[18] This completes the UE-specific processing.

Subsequently, the data transmitted to different UEs is multiplexed (step Ⓕ). Thereto, the precoded symbols are inserted into a common data structure which encompasses the whole subframe. This data structure also integrates other physical channels (e. g. PDCCH and PBCH) and signals (e. g. CRSs and CSI-RSs) into the radio frame.

The following processing steps operate on the combined subframe. The signal for each transmit antenna is processed independently. First, the FFT is used to calculate the inverse DFT for each OFDM symbol (step Ⓖ). Then, the CP is prepended to each OFDM symbol and the symbols are concatenated to a continuous signal (step Ⓗ). The resulting signal is then filtered digitally (step Ⓘ). For example, *digital pre-distortion* (DPD) anticipates and compensates non-linear effects of the power amplifier. Finally, the signals are converted to the analog domain, where they are further filtered, amplified, and forwarded to the antennas.

The receiving side of the eNodeB is not shown in the figure. In principle, that performs the same tasks in reverse. The most prominent difference is that DFT-spread OFDM requires a second DFT operation per UE. In addition, MIMO processing and demodulation have to estimate the UL radio channel from the embedded reference signals.

To make efficient use of the radio channels, LTE performs LA. Thereto, the eNodeB uses MIMO modes which are best suitable for the current channel conditions. In addition, it applies AMC, i. e. it selects modulation schemes and code rates to balance capacity and robustness. These are mechanisms which operate independently for each UE. The eNodeB furthermore applies opportunistic scheduling to allocate only those parts of the radio spectrum to each UE which can be used efficiently. All these mechanisms rely on CSI, which has to be acquired from measurements performed by the UEs.

Certain timing-constraints apply to the shown processing chain. First, the sizes of the transport blocks depend on the amount of radio spectrum allocated to each UE. Thus, the processing can only begin after the RA has completed its allocations. Second, the output of the processing chain has to be delivered to the DAC in time, so that the specification of the radio interface is met. Further restrictions on the timing are imposed by the design of the HARQ protocol. These imply that, to make best use of the available HARQ processes, after receiving an UL subframe the eNodeB has to perform RA and the DL processing in about 3 ms [DPS16, section 8.3.1].

After RA is completed, the user specific processing can be parallelized easily, because there are no interdependencies between different UEs. Furthermore, when a subframe has been assembled, the further processing can be performed independently for each transmit antenna. In addition to these options for parallelization, the described timing constraints suggest a pipelined implementation. For example, the cell specific processing for subframe $n$, the user specific processing for subframe $n + 1$, and the RA for subframe $n + 2$ could be processed simultaneously.

[18] These operations are depicted in more detail in figure 2.7. Note that, in contrast to that figure, CRSs are here shown to be integrated during frame assembly, because they are not associated with a single UE. Also, the mapping from antenna ports to antennas is not shown here.

## 2.4 Framework for the Management of Radio Resources

LTE systems can use a large bandwidth to achieve high throughput. The applied modulation scheme OFDM brings high flexibility, as the system bandwidth is split into many orthogonal subcarriers. These can be used for data transmission. In addition, reference signals have to be transmitted to support highly efficient modulation and coding and complex MIMO schemes. A sophisticated resource management is required to achieve efficient utilization of the resources.

In this thesis, the discussion of the management and allocation of radio resources is split into two parts. This section introduces the framework for the allocation of channel resources that is standardized in LTE. The actual algorithms that decide which resources are allocated to which UE are not covered by the LTE standardization. These algorithms are introduced in chapter 3.

The introduction of the framework for resource management is structured into three subjects. First, section 2.4.1 defines how the channel resources are divided into orthogonal units. Second, section 2.4.2 describes how the different channels and signals used by LTE are mapped to these units. Third, section 2.4.3 targets the signaling mechanisms used to inform UEs about their allocated resources. Finally, section 2.4.4 provides a summary and discusses implications of this framework on resource allocation.

### 2.4.1 Structuring of Channel Resources

This section describes the general structure of the radio resources available to the LTE air interface. Section 2.4.1.1 covers the coarse-grained structures of carriers for UL and DL transmission. Subsequently, section 2.4.1.2 introduces the resource grid which resembles the fine-grained subdivision of a single carrier.

#### *2.4.1.1 Duplexing Schemes and Carriers*

LTE supports different frequency bands. The list of bands, which is specified in [3GPP 36.104], is continuously extended to keep up with global frequency allocations. Each band has a duplexing scheme assigned, which is introduced in the following paragraphs.

The most efficient duplex scheme is *full duplex*, where each of two communication partners transmits and receives at the same time on the same resources. However, when information is transmitted via a wireless channel, there is a significant imbalance between the transmitted power and the power of the received signal. In full duplex operation, this would cause strong self-interference. Therefore, wireless systems do typically not employ full duplex.[19] Instead, transmission and reception of signals are separated in time and / or frequency.

For maximum flexibility LTE supports two duplex schemes: In *frequency division duplex* (FDD), UL and DL transmissions are performed on two separate carriers. Both carriers are not directly adjacent, but separated by a guard bandwidth. This allows a receiver to filter out the band on

[19] Full duplex is in principle possible if self-interference can be canceled. However, this is still a research topic and not implemented in real systems [Zha+16]

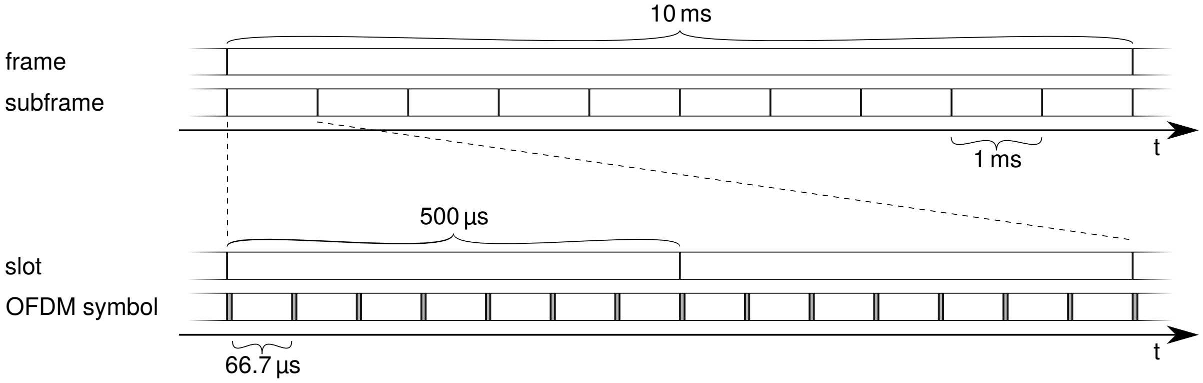


**Figure 2.9:** Time domain structure of the LTE radio frame for a FDD system with normal CP

which the local transmitter operates and thereby suppress self-interference. For FDD, two carriers have to be allocated, which is sometimes termed *paired spectrum.* In *time division duplex* (TDD), a single carrier is alternately used for UL and DL transmissions. Guard times between the two sections are required to switch from transmit to receive mode and vice-versa. These guard times also have to account for propagation delays. In Germany and Europe FDD systems predominate the marked [BNetzA16], while TDD is more common in Asia. This thesis focuses on FDD.

LTE supports different carrier bandwidths, so that it can be deployed flexibly. According to [3GPP 36.104], the bandwidth of a carrier can be 1.4 MHz, 3 MHz, 5 MHz, 10 MHz, 15 MHz or 20 MHz. The subcarriers have a constant bandwidth of 15 kHz, but the number of subcarriers depends on the carrier bandwidth. The transmitted signals do not cover the full bandwidth, but the numbers already include small guard bands, which are left free.

To support wider bandwidths while being compatible to previous releases, a single BS can host multiple cells, each offering service on a different carrier. However, in Releases 8 and 9, a UE can be connected only to a single cell at a time. Therefore, the peak data rate per UE is limited by the bandwidth of a single carrier. To mitigate this drawback, *carrier aggregation* (CA) is introduced by Release 10 [3GPP 36.300]. This allows a UE to be simultaneously connected to and served by multiple cells. Aggregated carriers have independent PHY layers. The combination of the carriers is performed by the MAC layer, so that it is invisible to higher layers [DPS16].

#### *2.4.1.2 Definition of a Resource Grid*

In [3GPP 36.211], LTE defines a two-dimensional grid of resources. In time domain, the largest unit is a radio frame with a duration of $307200 \cdot T_s = 10\,\text{ms}$.[20] The structure of radio frames is different for TDD and FDD systems, and is here presented only for the latter. A radio frame consists of ten subframes with a duration of 1 ms each. The subframe is the scheduling interval, which is also known as *transmission time interval* (TTI). Each subframe comprises two slots of duration $15360 \cdot T_s = 0.5\,\text{ms}$. A slot comprises either seven or six OFDM symbols, depending on the length of the CP. This structure is depicted in figure 2.9. It is valid for UL and DL operation.

In frequency domain, the resource grid consists of subcarriers. The number of subcarriers depends on the carrier bandwidth. For example, a system with 20 MHz uses a DFT with 2048

[20] With $T_s \approx 32.6\,\text{ns}$ as introduced in section 2.3.2.3.

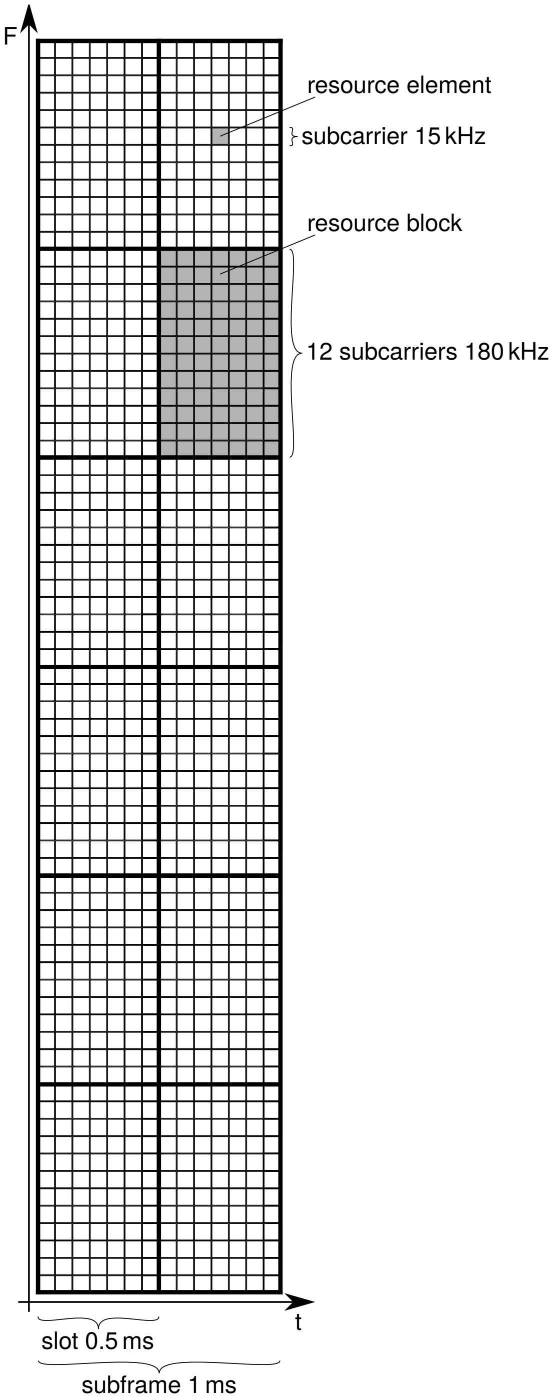


**Figure 2.10:** Resource grid for one subframe and 1.5 MHz carrier bandwidth

points. However, only 1200 subcarriers are used to transmit data, while the others are left free as guard band or to avoid effects with the central carrier.

The resulting two-dimensional resource grid consists of *resource elements* (REs). Each RE is a subcarrier in a single OFDM symbol, and carries a modulation symbol as introduced in section 2.3.2.1. To simplify the mapping of resource elements to transmitted data, LTE also defines *physical resource blocks* (PRBs). A PRB consists of the REs of twelve consecutive subcarriers and seven consecutive OFDM symbols, thus 84 REs. The resource grid for one subframe in a system with the minimum bandwidth of 1.5 MHz is depicted in figure 2.10.

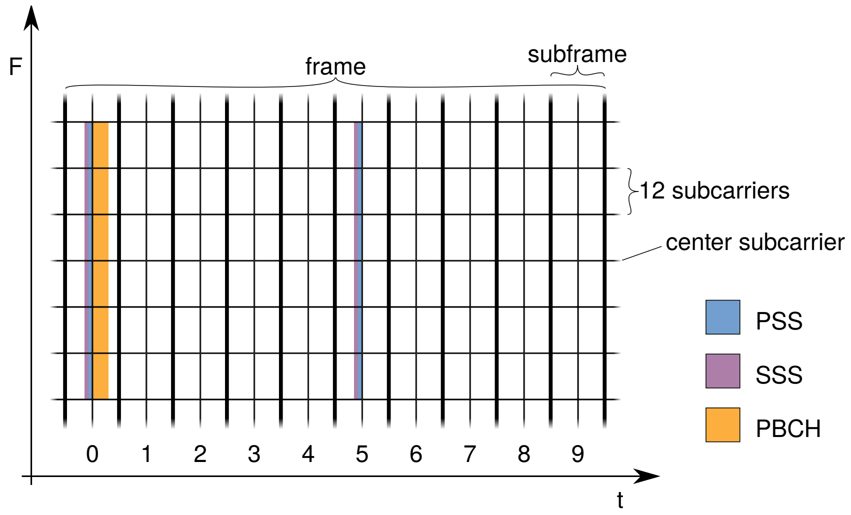


**Figure 2.11:** Placement of synchronization signals and PBCH

## 2.4.2 Mapping of Physical Channels and Signals to Resources

As introduced in section 2.2.1, LTE uses physical channels to describe the usage of resource elements. Physical channels either have a transport channel assigned, i. e. carry information from higher layers, or carry PHY layer control information. In addition, physical signals are transmitted. The mapping of channels and signals to REs is discussed in the following two sections for DL and UL, respectively.

### *2.4.2.1 Downlink*

In DL direction, an eNodeB transmits multiple physical channels as listed in section 2.2.1. In addition, it transmits synchronization signals and reference signals. Synchronization signals are used by UEs to detect a cell and synchronize on the DL carrier. Reference signals serve to estimate the DL channels and to decode transmissions on the physical channels.

The *Primary Synchronization Signal* (PSS) and the *Secondary Synchronization Signal* (SSS) allow a UE to learn the frame timing of the cell, the used duplex mode, and the cell's *physical-layer cell id*. For FDD, they are transmitted in subframes 0 and 5 of each frame. There, PSS and SSS occupy the central 72 subcarriers in the sixth and seventh OFDM symbols, respectively. Basic information required by UEs to connect to a cell is broadcasted on the PBCH. This occupies the central 72 subcarriers in the first four OFDM symbols of the second slot in subframe 0 of each frame. The location of the REs for PSS, SSS, and PBCH is marked in figure 2.11. These REs are not used for other transmissions.

The location of reference signals is defined per subframe. It is shown for two consecutive PRBs of a subframe in figure 2.12.

CRSs are transmitted on one, two, or four antenna ports. To avoid interference between antenna ports, a RE used to transmit a reference symbol on one of the antenna ports is left free on the other antenna ports. Each cell transmits one of 504 different sequences of reference symbols, corresponding to its physical-layer cell id. Also based on the physical-layer cell id, a shift in

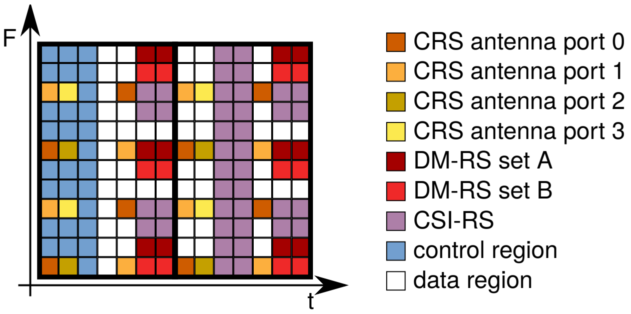


**Figure 2.12:** Allocation of REs to physical channels and signals in DL

frequency of up to five REs is applied to the positions of the CRSs. The power of the reference symbols can be boosted compared to the power of the other symbols. These approaches allow reliable estimation of the channel even in the case of strong interference. In total, four, eight, or twelve REs are occupied by CRSs when one, two, or four antenna ports are configured, respectively.[21] These are always transmitted, even if DM-RSs are used for demodulation and CSI-RSs for channel estimation.

DM-RSs can be used instead of CRSs for demodulation (see also section 2.3.4.4). Two different structures of DM-RSs were standardized for one antenna port in Release 8 (TM 7) and for one and two antenna ports in Release 9 (TM 8). The latter was later extended for up to eight layer spatial multiplexing (TM 9 and TM 10) and is presented here. Two disjoint sets of REs can be used for DM-RSs (here denoted as A and B; see also figure 2.12). These and orthogonal sequences of reference symbols are used to separate the antenna ports. The transmitted sequences of reference symbols depend on the cell ID or, since Release 11, on a per-UE configuration. DM-RSs are only transmitted on those PRBs allocated to a UE that use TM which relies on DM-RSs. Therefore, they do not impact transmissions to other UEs. A similar structure of reference symbols is also used for demodulation of EPDCCH.

CSI-RSs can be used to measure the DL channel as described in section 2.3.6.3. The REs which can be used for CSI-RSs are colored purple in figure 2.12. The periodicity and position of the CSI-RSs can be configured.

The remainder of the subframe is divided into a control and a data region. The control region carries the PCFICH, the PDCCH, and the PHICH. The data region carries all other physical channels. Despite being a control channel, the EPDCCH is transmitted in the data region. To dynamically reduce overhead, the control region has variable size and covers one to three OFDM symbols.[22] Its size is signaled via the PCFICH in the first OFDM symbol of each subframe.

The PDSCH is used to serve multiple UEs simultaneously. The allocation of these resources to UEs is flexible. It can be adapted to the channel conditions and the higher layer requirements. The objectives and strategies for this allocation are discussed in chapter 3.

[21] The density of CRSs on antenna ports 0 and 1 is higher than on ports 2 and 3. This results in a more robust channel estimation on these ports.

[22] Or two to four for small bandwidths.

#### *2.4.2.2 Uplink*

In UL direction, UEs transmit physical channels as defined in section 2.2.1. In addition, two types of reference signals are transmitted. DM-RSs aid the decoding of PUCCH and PUSCH. SRSs are transmitted to allow the eNodeB to measure the UL channel (see section 2.3.6.2).

A low PAPR is important to avoid overly complex power amplifiers in the UEs (see also section 2.3.2.2). Therefore, the modulation of each channel and signal is designed to minimize PAPR. In addition, frequency multiplexing of channels and signals of a single UE is avoided. Instead, transmissions are multiplexed in the time domain. Before Release 10, a UE never transmits on PUCCH and on PUSCH simultaneously. Also, the RA mechanism ensures allocation of adjacent subcarriers in PUSCH. This restriction is partially relaxed in Release 10 for UEs whose power amplifiers have sufficient headroom.

In the PUSCH, DM-RSs are transmitted on all subcarriers in the fourth OFDM symbol of each slot. The sequence of reference symbols depends on the physical-layer cell ID or on terminal specific configuration. It has been designed to achieve orthogonality between spatial streams transmitted by the same UE and between multiple UEs in the same cell transmitting on the same PRBs in a MU-MIMO system. In addition, UEs of neighboring cells should also use pseudo-orthogonal sequences.

The PUCCH uses a different method to transmit reference symbols for demodulation, which is also based on orthogonal sequences. To avoid segmentation of available bandwidth, PUCCH resources are allocated from the lowest and highest PRBs of a carrier. To increase robustness, frequency hopping is always applied for PUCCH.

SRSs are transmitted with configurable bandwidth on every second subcarrier in the last OFDM symbol of a subframe. UEs are configured to do not transmit on the respective OFDM symbol whenever that is used for SRSs by the same or another UE.

### 2.4.3 Signaling of Allocated Resources

Every subframe, an eNodeB performs RA for PDSCH and PUSCH. The resource grants for UL and DL are then communicated to the UEs via the PDCCH. These PHY layer control messages also contain information about modulation scheme, code rate, HARQ, MIMO, and power control. The specification of the allocated resources in the signaling message imposes certain restrictions on the RA. Also, the capacity of the control channel is limited, which can prohibit signaling of combinations of sets of allocated resources. These topics are discussed in the following sections.

#### *2.4.3.1 Specification of Allocated Resources*

Resources of PDSCH and PUSCH are allocated with a granularity of pairs of PRBs. Each PRB pair is transmitted in consecutive slots of the same subframe. To support frequency selective and frequency diverse allocation of resources, LTE introduces *virtual resource blocks* (VRBs) as an abstraction for resource specification [3GPP 36.211]. The eNodeB allocates pairs of VRBs, which are then mapped to pairs of PRBs by different mapping functions. The type of mapping used is signaled to the UE as part of the RA grant.

To allow for frequency selective scheduling, VRBs can be mapped directly to PRBs, so that adjacent VRBs are mapped to adjacent PRBs (*localized mapping*). For DL LTE also supports *distributed mapping* for a frequency-diverse RA. There, adjacent VRBs are mapped to distant PRBs. In addition, the mapping is different in both slots of a subframe. A similar concept is implemented by frequency hopping in UL.

After allocating resources for DL transmissions, the eNodeB has to encode the allocation to communicate it to the UEs. For the design of this encoding, diverse objectives were taken into account. The encoding has to be efficient, while providing flexibility for frequency selective scheduling when required. At the same time, it has to support allocations consisting of a single up to all VRBs. Therefore, three different RA types are defined in [3GPP 36.213].

For type 0, VRBs are divided into *resource block groups* (RBGs), such that each RBG consists of consecutive VRBs. The group size depends on system bandwidth.[23] The eNodeB can then allocate arbitrary groups to a UEs by specifying a bitmask. Type 0 is only used with localized VRB mapping. Type 1 uses the same definition of RBGs as type 0. Here, the RBGs are aggregated to sets, where the number of sets equals the RBG size. An interleaving pattern allocates the RBGs to the sets. The eNodeB specifies one of the sets and a bitmask to select one or multiple VRBs out of that set.[24] As type 0, type 1 is only used with localized VRB mapping. In contrast to the previous RA types, type 2 is used for the allocation of a continuous range of VRBs with low overhead. This is the only RA type also supporting distributed VRB mapping.

The three RA types target different objectives. Types 0 and 1 are optimized for frequency selective scheduling. While type 0 allows to allocate arbitrary VRBs, the granularity is restricted to full RBGs. Type 1 allows the scheduler to allocate single VRBs, but restrictions apply for the combination of VRBs. Type 2 imposes less overhead, but further restricts the combination of VRBs. It is preferred for frequency diverse scheduling. Thus, the RA types are applicable for different classes of UEs.

A similar approach is used to encode UL allocations. It is simpler than the scheme for DL, because the UL allocations have to be continuous to achieve a low PAPR. Two RA types are defined for UL. Type 0 is equal to type 2 as specified for DL. Instead of distributed VRB mapping, it allows to enable frequency hopping. Type 1 is introduced in Release 10. It is based on the definition of RBGs as introduced for DL. It allows to signal two blocks of consecutive RBGs. Here, frequency hopping not supported.

#### *2.4.3.2 Transmission of Resource Allocation Messages*

After being encoded with one of the types described above, RA messages have to be transmitted to the UEs via PDCCH or EPDCCH. The capacity of these control channels is limited. The following paragraphs consider the capacity of the PDCCH.

[23] The size of the RBGs is one, two, three, and four for systems with up to 10, 26, 63, and 110 PRBs, respectively.

[24] To reduce size of the allocation bitmask, that covers only a part of a set. A single bit encodes which part.

As introduced in section 2.4.2.1, the size of the control region of a subframe is adapted dynamically. It is shared by PCFICH, PHICH, and PDCCH. The size of the control region is communicated via the PCFICH, which encodes these two bits into 16 REs. The PHICH carries HARQ feedback for UL transmissions. The number of REs required for PHICH is semi-statically configurable, as it has to be adapted to the number of UL RAs. The remaining capacity of the control region is available for PDCCH.

The PDCCH carries scheduling assignments for UL and DL allocations and power control commands, together termed *downlink control information* (DCI). While allocations for DL are valid for the same subframe, allocations for UL are transmitted four subframes in advance. Multiple DCIs are transmitted in the same subframe. A single DCI can be intended for multiple UEs. This is used for power control or transmission of system information. Also, a single UE can receive multiple DCIs in the same subframe.

In addition to the allocated resources, other information has to be signaled to the UE. This information depends on the TM. To use the available channel capacity as efficient as possible, 3GPP defines multiple formats to encode the DCI [3GPP 36.212]. The formats have different sizes. In addition, LA adapts the transmit power and the FEC encoding to the radio channel conditions. Together, this results in a variable number of occupied REs per DCI.

The allocation of REs to DCIs is based on *control channel elements* (CCEs). A CCE is a group of 36 REs with a fixed location. Depending on its size and coding, a DCI occupies one to eight CCEs. The number and identity of CCEs intended for a UE are not known to that UE. Instead, each UE has to blindly try to decode multiple possible configurations. To simplify the decoding attempts, a combination of common and terminal-specific search spaces are defined. All UEs in a cell try to decode CCEs in the common search space. In addition, each UE tries to decode CCEs in its terminal-specific search space. The terminal-specific search spaces are defined by terminal identity and subframe number. For efficient utilization of the resources, the terminal-specific search spaced do overlap.

To supplement the PDCCH, the EPDCCH is introduced in LTE Release 11. It is transmitted in the data region of a subframe on PRBs selected by the eNodeB. This allows for frequency selective scheduling and interference coordination (see section 4.2). The EPDCCH is decoded using DM-RSs, so that advanced MIMO mechanisms can also be used for the control channel. The eNodeB can dynamically decide to transmit EPDCCH or use the same PRBs for the PDSCH.

Possible locations of the EPDCCH are configured semi-statically and separately for each UE. Each UE monitors one to two EPDCCH sets. Each set consists of two, four, or eight PRB pairs located at arbitrary positions in the frequency domain. EPDCCH sets can be configured to be localized or distributed, resulting in adjacent or distributed REs allocated to a single DCI, respectively. Otherwise, the allocation of REs to DCIs works similar as for PDCCH. To decode the EPDCCH, UEs also use overlapping search spaces.

Even with efficient encoding, transmission of RA DCIs imposes a significant overhead for small data messages. Therefore, in addition to dynamic RA, LTE allows for semi-persistent RAs. Without further signaling, the RA then repeats with an interval configured by higher layers. Semi-persistent scheduling has been designed to efficiently handle VoIP calls, but can also be used for other communication.

In case of a single carrier system, RAs signaled on PDCCH and EPDCCH implicitly relate to the PDSCH on the same DL carrier or the PUSCH on the cell's UL carrier. Systems that apply CA can operate in the same manner. The required association of UL to DL carriers is configured statically. In addition, LTE allows for cross-carrier scheduling. This is configured per UE and per carrier by higher layer configuration. For each UL and DL carrier, the system configures an associated signaling carrier, whose PDCCH transmits the RAs. Control information then has to be augmented with a carrier indicator.

### 2.4.4 Summary and Discussion of Impacts on Resource Allocation

In principle, PDSCH resources can be freely assigned to UEs for DL transmissions. For UL direction, resources have to be continuous, but some flexibility is left. The RA mechanism can use this to optimize system efficiency in both directions. However, it has to consider various restrictions, which stem from the mapping of channels and signals, limited specification options, and restricted control channel capacity for signaling.

In DL, synchronization signals reduce the capacity of some PRBs. Similarly, CSI-RSs reduce the capacity for those UEs which know about them. In contrast, these reference signals increase the error rate for legacy UEs. In UL direction, the capacity of the PRBs is only impacted by SRSs.

To avoid overhead for the signaling of RA messages, efficient encoding options have been defined. However, for the DL direction these also impose constraints on RA. Depending on the chosen RA type, either the granularity of RA is confined, or the possible combinations of resources are limited. In addition, the RA types used for different UEs have to be carefully combined to allow orthogonal allocation of all resources. For UL direction, most restrictions come from the continuous allocation of resources.

The capacity of the control channels, which have to be used to inform UEs about RAs, is limited. The resources required per DCI depend on various parameters. They are influenced by the configured TMs and the applied RA type. As search spaces of different UEs overlap, allocations of CCEs can conflict and render signaling to a UE impossible. The size of the PDCCH can be configured in a limited range. However, a larger PDCCH comes at the cost of reduced capacity of the data region. A similar trade-off also applies to the EPDCCH. While semi-persistent scheduling reduces the load on the control channels, it imposes a restriction on RA in the following subframes.

UL and DL RA mechanisms compete for control channel resources and have to meet the described restrictions. Some decisions, like selection of the RA type or configuration of the size of the control region, can be taken every subframe. Others, like TM, EPDCCH sets, and semi-persistent scheduling, need to be configured in advance. Monitoring and optimization of the configurations therefore has to be performed on a larger time interval. All these constraints complicate RA.

# 3 Management of Radio Resources

*Scheduling* in general describes the allocation of resources to jobs, typically by determining the order in which the jobs access the resources. In the context of networking, determining an order in which packets are sent over a link is termed *packet scheduling*. The capacity of mobile channels varies over time and frequency. To cope with this, the system provides flexible options to allocate time and bandwidth resources to different users. This allows for *opportunistic RA* to maximize system performance. Note that in the topic of mobile networks the terms scheduling and *resource allocation* are used synonymously. This thesis favors the term resource allocation, because that emphasizes that the mechanisms are not only related to a timely order of data transmissions.

LTE does not specify RA mechanisms, but defines a framework where the eNodeB vendors can plug in their own mechanisms. Section 3.1 gives an overview over this framework and other constraints for RA originating from the LTE specification. RA provides many degrees of freedom, which can be used to maximize different network performance metrics and other criteria. Different objectives for RA are described in section 3.2.

RA can be seen as optimization problem. Multiple formulations are presented in section 3.3. The optimization problems are often too complex to be solved in each BSs. However, they are used to gain insights into characteristics of the problem or to serve as benchmark for system evaluation. Heuristics are then implemented in real systems, which strive to achieve high performance with manageable computational complexity. Such heuristics are reviewed in section 3.4. The chapter is concluded by a discussion in section 3.5.

## 3.1 Integration of Resource Allocation into the LTE System

RA in LTE is embedded into a larger bandwidth management system. The main functional units of this are depicted in figure 3.1. The *admission control* (AC) decides whether a new bearer can be accepted. This helps to prohibit overload and impairment of service quality for the new and the previously existing bearers. A rate shaping unit limits the rate of one or multiple flows of data. It thereby ensures that flows comply with the subscription of a user. In case a flow exceeds its rate limit, the rate shaping unit may also queue and drop packets.

Before being transmitted over the wireless link, packets are queued in the eNodeB. In case the wireless link is permanently congested, packets are dropped from that queue. This is required to cope with finite buffer capacity. It also avoids to transmit packets which are delayed for too

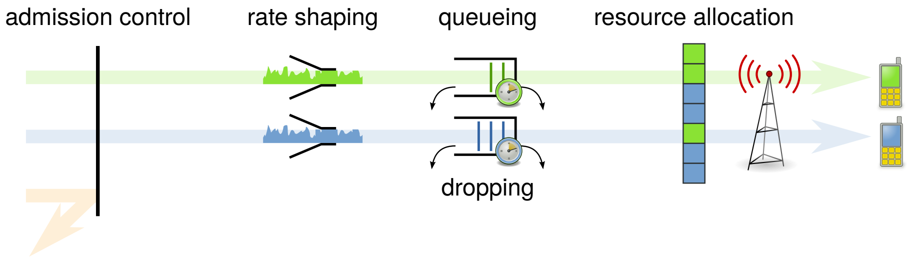


**Figure 3.1:** Overview over bandwidth management and RA mechanisms

long and are therefore not useful to the receiver any more. Transport protocols get notified of the congestion either by packet drops or by *explicit congestion notification* (ECN).[1] In the eNodeB, time and bandwidth resources (also termed *air interface resources*) are allocated to bearers. This is performed in collaboration with LA mechanisms.

Each of the functions of the bandwidth management system influences the service quality of the network. We do here focus on the allocation of time and bandwidth resources.

RA algorithms for UL and DL operate independently in the eNodeB at the 1 ms TTI. They allocate resources in the PDSCH and PUSCH, respectively. Typical input data for RA are queue states (length, age of packets), channel information, the requested service quality per bearer, and measurements of previous service quality (e. g. bit rate, packet loss ratio) [Cap+13].

As main output, time and bandwidth resources are allocated to bearers (or UEs for UL direction). In addition, the transmit power can also be interpreted as an allocated resource. Depending on the modulation scheme and the TM, it can be configured dynamically per TTI or semi-statically. Related to these allocations is also the determination of LA parameters (including modulation, coding, and MIMO configuration). If supported by the system, the selection of MU-MIMO sets can also be seen part of the RA. MU-MIMO is not investigated in this thesis. In addition to radio related resources, in this thesis also compute resources required for the processing of the transmissions are allocated. However, allocation of such resources is not discussed in this chapter, but reviewed in section 5.2.

The LTE standards impose different constraints on the RA. Protocols such as HARQ specify timing constraints, which have to be followed. Also, there are other aspects such as control signaling and synchronization, which require transmission of messages in limited time. The flexibility of specification and signaling of allocated resources is limited as described in section 2.4.3. Finally, there are physical limits such as power constraints.

## 3.2 Objectives

The main objective for RA is to maximize the service quality of all users. As the users compete for the shared resources, the RA has to balance their demands. In the following section, models and metrics for the service quality from a single user perspective are introduced. Based on that, section 3.2.2 discusses how competing demands of different users can be combined to an objective from network perspective.

[1] ECN is introduced into LTE by Release 9 [3GPP 36.300 v9.1.0].

### 3.2.1 Objectives from a Single User Perspective

The objectives of a single user can be differentiated into those related to the performance of the network service and other objectives. Other objectives, such as power consumption at the UE, are not discussed here.[2] The main objective for RA from the users' perspective is the quality of the service provided by the network (QoS). RA is not the only, but one of the major influencing factors of QoS.

A consistent terminology for QoS is introduced by Gozdecki, Jajszczyk, and Stankiewicz [GJS03]. They differentiate between three notions of QoS. The *intrinsic* QoS is the objectively measurable performance of the network. This is called *network performance* by ITU and *European Telecommunications Standards Institute* (ETSI), and QoS by *Internet Engineering Task Force* (IETF). The *perceived* QoS represents the subjective service quality as perceived by the user, which partially depends on the intrinsic QoS. The *assessed* QoS resembles the decisions of a user to use a service. The latter also depend on pricing and customer service. This thesis is concerned with the intrinsic performance of the network.

Before QoS can be discussed, a model for the service offered by the network has to be defined. The simplest model for a data transmission in a network is that of fluid flows. Multiple continuous streams of data with finite rate are served simultaneously and share the resources of the network [Ada97]. Resources can be assigned to the flows with arbitrary granularity. The rate of a flow is either limited by the source or by the resources available in the network.

A more commonly used and more realistic model is that of a packet-based service. A packet is a block of data with finite length. It is transmitted from sender to destination in non-zero time. The packet can be queued or dropped by the network. A source inserts packets into the network with a constant or variable rate, typically without taking into account the load of the network.

A third model for the service offered by a network is that of objects on application layer [Pro+12; Pro15; Kas16]. Transmission of objects models the transmission of *hypertext markup language* (HTML) pages, images, e-mails, etc. with the help of application and transport protocols. Objects have finite size. While multiple objects can share the network resources following a fluid model or by being segmented into packets, they are re-aggregated at the receiver and delivered as one piece to the destination. Like the fluid flow model, this model reflects the elasticity of many Internet applications, i. e. their ability to adapt to limited network performance by increasing response times.

Based on the service model, QoS metrics can be defined. Gozdecki, Jajszczyk, and Stankiewicz [GJS03] introduce four QoS metrics (called parameters there): bit rate, delay, jitter (variance of delay), and packet loss ratio. The definitions of delay, jitter, and packet loss ratio are based on packets. The bit rate corresponds to the fluid flow model. However, this metric can also be derived from packet performance by averaging over certain time.

Some of the metrics are conflicting and can be traded for others without effort. For example, without modification of the RA, packet delay and jitter can be improved by dropping delayed packets. This illustrates that these metrics have to be evaluated jointly.

[2] Further non-performance related objectives are listed by Cao and Li [CL01] and Capozzi et al. [Cap+13].

Following the service model of objects on application layer, QoS can also be evaluated on higher layers [Pro+12; Pro15; Kas16]. Proebster [Pro15] and Kaschub [Kas16] define utility as a function of the time it takes to transmit one or a group of application layer objects, the size of the object(s), and a priority. When the application layer object is not transmitted within a certain time, a timeout occurs and the request is aborted. The time the network takes to transmit an object is a metric for the intrinsic service quality on application layer. The utility and the timeout model the expectations and behavior of the users, and thereby map the intrinsic service quality to the perceived service quality.

In contrast to metrics on packet level, object transmission times are better able to capture the performance as perceived by the user. Especially, this metric also captures interactions of the network with higher layer protocols. The widely used TCP, e. g., is highly sensitive to link-related packet drops.[3] This interaction can be captured by application layer metrics. A drawback of metrics on application layer is that they are not fully controlled by the network. As the transport protocol is implemented by the end systems, their configuration influences the metrics.

A user may use the network multiple times by transmitting multiple data flows, packets, or application layer objects. The intrinsic service quality can be evaluated independently per network usage, or aggregated metrics can be derived. For a packet based network model, the data rate is an aggregated metric which is calculated by averaging the transmitted data volume over time. Also, the jitter is an aggregated metric based on the variance of packet delays. Other typical means of aggregation are averaging and deriving of extreme values. An example is the delay bound, which describes a delay value that is never exceeded.

3GPP states performance requirements for LTE in [3GPP 36.913]. These encompass the theoretical peak data rate, the control plane latency (time to setup a connection), and the user plane latency (delay of the first packet). These metrics focus on static system parameters and signaling procedures, but not on RA.

In addition, 3GPP specifies a QoS framework [3GPP 23.401; 3GPP 23.203], which is based on bearers as introduced in section 2.2.1. Each bearer is assigned a *QoS class identifier* (QCI), a scalar indicating certain QoS requirements. The standardized QCIs and associated performance criteria are listed in table 3.1. The QCI assigns one of two resource types to each bearer: Either *guaranteed bitrate* (GBR) or *Non-GBR*.

GBR bearers have the parameters GBR and *maximum bitrate* (MBR). These can be interpreted as follows. For bearers not exceeding the GBR, no congestion may occur and the system has to meet the maximum delay and packet loss ratios associated with the respective QCI. In case the data rate of such a bearer exceeds the MBR, it can be subject to rate shaping. While for Releases 8 and 9, GBR and MBR were set to equal values, since Release 10 the MBR can be larger than the GBR.

For Non-GBR bearers, an average MBR can be specified which limits the sum data rate of multiple bearers. In case that rate is exceeded, the rate is shaped, which involves delaying and dropping of packets. Otherwise, packets are transmitted as possible. The system should not exceed the packet delay and the packet loss ratio that are given in table 3.1. However, packet loss in the table only relates to losses on the radio link, not encompassing congestion related drops.

[3] In contrast, congestion related packet drops are required to control the transmit rate.

**Table 3.1:** Characteristics for service quality on the radio interface standardized by 3GPP (adapted from [3GPP 23.203 v10.10.0, table 6.7.1]). In the priority column, small numbers indicate high priority.

| QCI | Type | Prio. | Radio Link Packet Delay [ms] | Radio Link Packet Drop Probability | Services (exemplary) |
|---|---|---|---|---|---|
| 1 | GBR | 2 | 80 | $10^{-2}$ | Conversational Voice |
| 2 | GBR | 4 | 130 | $10^{-3}$ | Conversational Video (Live Streaming) |
| 3 | GBR | 3 | 30 | $10^{-3}$ | Real-Time Gaming |
| 4 | GBR | 5 | 280 | $10^{-6}$ | Non-Conversational Video (Buffered Streaming) |
| 5 | Non-GBR | 1 | 80 | $10^{-6}$ | Signaling, e. g. for VoIP |
| 6 | Non-GBR | 6 | 280 | $10^{-6}$ | TCP-based, Video (Buffered Streaming) |
| 7 | Non-GBR | 7 | 80 | $10^{-3}$ | Voice, Video (Live Streaming),<br>Interactive Gaming |
| 8 | Non-GBR | 8 | 280 | $10^{-6}$ | TCP-based, Video (Buffered Streaming) |
| 9 | Non-GBR | 9 | 280 | $10^{-6}$ | TCP-based, Video (Buffered Streaming) |

This allows the system to serve a bearer with an arbitrary low data rate. Consequently, following these definitions is not sufficient to deliver a good service to the users.

The QCI also assigns a priority to the bearers. This is not necessarily higher for bearers with the GBR resource type. The priority defines the order in which requirements of bearers competing for insufficient resources shall be violated.

Setting up bearers for specific services has to be supported by the network operator and by the UE. GBR bearers are used for voice and video calls that are offered as a service by the network operator. Most of the Internet traffic, including voice and video not served by the network operator, is handled in a default bearer, e. g. with QCI 8.

The LTE QoS framework defines some metrics and gives guidelines. However, this framework is not sufficient to specify the service quality that should be achieved by a RA mechanism. Especially for Internet traffic transmitted via TCP, additional metrics have to be considered.

### 3.2.2 Objectives from Network Perspective

The main objective of a mobile network operator should be to satisfy as many customers as possible with an economically efficient system. Intuitively, satisfying customers can be achieved by transmitting as much of the user's requested data as quickly as possible. This translates to the maximization of the sum or average throughput. An equivalent metric is the *spectral efficiency*, which normalizes the throughput by dividing it by the system bandwidth [SAR09]. As often the revenue of an operator depends on the data volume transmitted by its customers, maximizing the total throughput is also attractive from an economical perspective.

In addition, the distribution of QoS between the users is also an important aspect. This is termed the *fairness* of the system. It can be defined as the allocation of scarce resources to competing users, so that every user experiences an acceptable service quality. In mobile networks, fairness and sum throughput are typically contradicting goals. This is caused by differences in channel quality. Allocating resources to users with low channel quality impairs the total throughput.

Achieving fairness can also be interpreted as maximization of the perceived QoS of all customers. The mapping from an intrinsic metric to perceived QoS is typically not linear. For example, lowering intrinsic QoS finally results in services (applications) becoming unusable and the customers being annoyed. At the same time, certain high-performance intrinsic QoS might be sufficient for a service, e. g. when more promptly reaction of an application cannot be noticed by the user.[4] As result from the non-linearity of the mapping, maximizing the sum of an intrinsic QoS metric is not equivalent to maximizing the perceived QoS. Instead, a more balanced intrinsic QoS typically leads to a higher perceived QoS.

Fairness can be defined either qualitatively, i. e. by stating criteria which have to be met for a system to be fair, or quantitatively, i. e. by defining a metric to measure the degree of fairness. Fairness can be evaluated on different time scales [AAR12]. Short-term fairness requires the system to be equally fair at all time, while long-term fairness allows temporary deviations as long as they are balanced out later. Fairness can be defined on different QoS metrics, e. g. a system can transmit fair data rates or guarantee a maximum packet delay for all users. A system can also be fair in the sense that the same amount of resources (e. g. bandwidth, power) are allocated to each user. In principle, the fairness scheme is independent of the parameters it is applied to [JCH84]. Here, fairness schemes are presented which are applied to the data rates of the users.

The most intuitive qualitative fairness criterion is that a system is fair when all users receive the same data rate. Mathematically, this results in maximization of the minimum rate of all users.[5] It is therefore called *max-min fair*. In some systems, not all users share the same bottleneck or RA granularity is not arbitrary. In those cases, the QoS of some users can be higher than the minimum, although the minimum cannot be further increased. Demers, Keshav, and Shenker [DKS89] also specify the QoS of those users. Their definition is based on a hypothetical system. In that, the user with minimum QoS and the resources occupied by this user have been removed. They state that a max-min fair RA also has to maximize the minimum QoS for this hypothetical system. In addition, this property has to hold recursively, i. e. when removing the user with the next-lowest QoS and its resources, the remaining system still has to be max-min fair.

Another qualitative definition of fairness is that of *Proportional Fair* (PF). It has been introduced by Kelly [Kel97] for fixed networks, but is also widely applied for wireless networks. A RA is proportionally fair if there is no other possible RA for which the sum of proportional changes of the data rates is positive. For the definition of the sum of the proportional changes, assume that

[4] Compare to the S-shaped utility function used by Proebster [Pro15] and Kaschub [Kas16].

[5] Given that the users compete for the same resources, this results in the same rate assigned to each user.

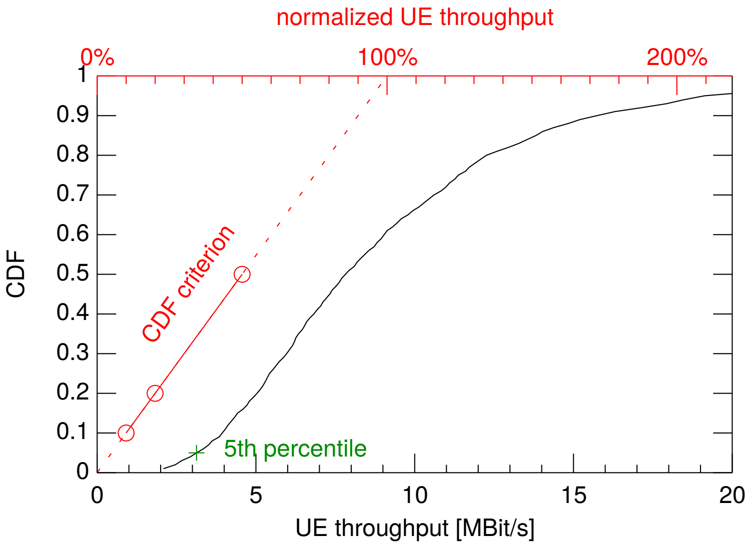


**Figure 3.2:** Exemplary CDF of UE rates with CDF-based fairness criterion and percentile

$\mathbf{x}$ is a vector of user rates with elements $x_i$, and $\mathbf{x}'$ is another vector of rates. Then, the sum of proportional changes from $\mathbf{x}$ to $\mathbf{x}'$ can be calculated as

$$C(\mathbf{x}, \mathbf{x}') = \sum_i \frac{x_i' - x_i}{x_i}. \tag{3.1}$$

A RA which is proportionally fair maximizes the sum of the logarithms of the users' rates.

Fairness can also be achieved by guaranteeing a fixed minimum rate to each user. The rate has to be chosen carefully, so that it is sufficient to provide useful service to the customers, but also not too high to not cost too many resources. Meeting a guaranteed rate can become impossible when the network is highly loaded or a user has a low channel quality.

As fairness describes the distribution of QoS between users of a system, it can also be evaluated by looking at the empirical *cumulative distribution function* (CDF) of a QoS metric. For each value $x$, an empirical CDF denotes the fraction of samples that were encountered with a measured value smaller than $x$. To be better able to describe fairness visible from a CDF, two CDF-based fairness definitions are commonly used for wireless networks.

The first, which is described by 3GPP2, *Next Generation Mobile Networks* (NGMN), and IEEE [NGM08; Sen+06; 3GPP2 04], is a qualitative criterion based on the normalized throughput. The normalized throughput is defined as the throughput of a user divided by the average throughput per user. A RA is considered to be fair if the CDF of the normalized throughput lies right of the line through the points $(0.1\,;0.1)$, $(0.2\,;0.2)$, and $(0.5\,;0.5)$. An exemplary empirical CDF of the UE rates in an LTE system is plotted in figure 3.2. The shown allocation is fair according to this criterion, because the CDF lies completely right of the red line.

The second CDF-based fairness metric is the widely used 5th percentile of the users' rates. It is marked by a green + in figure 3.2. When evaluated together with the average rate, it measures the performance discrimination experienced by a typical cell edge user. Users with even lower rate are ignored by this metric. This can be interpreted as being in outage, i. e. not being served by the network due to bad channel conditions.

Another popular metric for fairness is the fairness index introduced by Jain, Chiu, and Hawe [JCH84]. As above, assume that $\mathbf{x}$ is a vector of user rates with elements $x_i$ and length $n$. The fairness index $J(\mathbf{x})$ is then defined as

$$J(\mathbf{x}) = \frac{\left(\sum_i x_i\right)^2}{n \sum_i x_i^2} \tag{3.2}$$

For a RA which results in the same rates for all users, $J(\mathbf{x}) = 1$. For more unequal rate distributions the formula yields values between one and zero.

Providing fair and useful service to users with low channel quality or under high system load can be impossible or have a strong impact on the performance experienced by other users of the system. For some services, users may favor errors at session initialization over poor service quality during the session (e. g. voice calls, videos). AC can prevent the system from overload. RA and AC policies have to operate jointly to efficiently use the resources of a mobile network.

In addition to the objectives of sum rate and fairness, other aspects may be important for a RA mechanism. A comprehensive list of desired properties is given by Fattah and Leung [FL02]. In addition to QoS related aspects, implementation complexity and scalability of the RA algorithm are important requirements [CL01; FL02; SJT09; Cap+13].

### 3.2.3 Discussion

In an LTE network, users with optionally different objectives compete for resources. Contradicting objectives from network perspective also exist. The configuration of trade-offs, e. g. between fairness and system throughput, depends on policies of the network operator. Therefore, also in literature there is no single commonly accepted objective. Instead different publications present approaches that target different objectives. Also, many proposed algorithms provide parameters which allow the configuration of such trade-offs.

## 3.3 Resource Allocation as Optimization Problem

The allocation of resources to competing requests can be formulated as optimization problem [HL08]. In publications, this often serves to formulate the system model and the objective, before introducing realizable algorithms [And07]. Other authors, e. g. Lin, Shroff, and Srikant [LSS06], derive scheduling strategies and heuristics from analytical solutions of optimization problems.

### 3.3.1 Fixed Networks

Optimizing the order in which packets are served by a network is equivalent to the optimization of job scheduling, a prominent topic in operations research. A classification of problems and optimization approaches is given by Graham et al. [Gra+79]. Also in the context of fixed networks, but based on a fluid flow model, Kelly [Kel97] optimizes the routing of flows and the allocation of rates to flows that share multiple bottlenecks in a network. Basing on the introduction of utility by Shenker [She95], Kelly defines $\mathrm{U}(x)$ to be the utility resulting from rate $x$. This function can be different for each user of the network.

Assume that $\mathcal{J}$ is a set of resources in a network (e. g. links) and $c_j$ is the capacity of resource $j$. A route $r$ consists of the combination of one or multiple resources, i. e. $r \subset \mathcal{J}$ and $r \neq \emptyset$. The set of all possible routes is denoted as $\mathcal{R}$, and the subset of routes which share resource $j$ as $\mathcal{R}_j$, with $\mathcal{R}_j \subset \mathcal{R}$. Multiple routes can serve the traffic that originates from the same source and is destined to the same sink. A pair of source and sink is denoted as $s$, and the set of all such pairs as $\mathcal{S}$. The routes belonging to the same pair $s$ are grouped as set $\mathcal{R}_s$, with $\mathcal{R}_s \subset \mathcal{R}$. Each route belongs to exactly one pair, thus $\mathcal{R}_s \cap \mathcal{R}_t = \emptyset \; \forall (s,t) \in \mathcal{S}, s \neq t$.

The optimization problem adapts the allocation of rates to routes, and thereby to pairs of source and sink. The variable $y_r$, with $y_r \geq 0$, $r \in \mathcal{R}$, denotes the rate carried by route $r$. The problem is then defined as follows (from [Kel97], notation adapted):

$$\max_{y_r} \quad \sum_{s \in \mathcal{S}} \mathrm{U}_s\left( \sum_{r \in \mathcal{R}_s} y_r \right) \tag{3.3a}$$

$$\text{s. t.} \quad \sum_{r \in \mathcal{R}_j} y_r \leq c_j \qquad \forall j \in \mathcal{J} \tag{3.3b}$$

Here, the constraint (3.3b) guarantees that the routes sharing a resource do not exceed the capacity of that resource. The argument of the utility function in the objective (3.3a) is the rate allocated to a pair of source and sink $s$, which is here formulated as sum over the rates of all routes serving this pair.

The utility function can be defined so that the optimal RA has a desired fairness. Defining $\mathrm{U}(x) = x$ maximizes the sum throughput without considering fairness. Kelly also introduces a parametrized family of utility functions:

$$\mathrm{U}(x, \alpha) := -(-\log x)^{\alpha}, \qquad 0 < x < 1, \alpha \geq 1 \tag{3.4}$$

For $\alpha = 1$, equation (3.4) equals the logarithmic utility function, which results in a PF optimum. For larger $\alpha$, equation (3.4) gives higher priority to flows with low rates, and for $\alpha \rightarrow \infty$, the optimum approaches max-min fair.

### 3.3.2 Wireless Networks

The approaches developed for fixed networks can also be used for channel-unaware RA in mobile networks. However, the performance can be improved by taking the instantaneous channel quality into account.

Having this in mind, optimization problems have been defined for RA in mobile networks. Depending on the underlying network technology, resources in time, bandwidth, and / or power are allocated. Typically either fixed time slots or otherwise predefined resources are allocated, which is modeled by flag variables. In case the transmit power is adapted, the conversion of received power to data rate or capacity is part of the problem formulation. The objective can be either to maximize rates or utilities (rate adaptive RA) under limited transmit power, or to minimize transmit power while transmitting minimum rates (margin adaptive RA) [SAE05; SAR09]. This thesis focuses on the rate adaptive formulation.

A generic definition of the rate adaptive optimization problem, which is introduced in similar form by Shen, Andrews, and Evans [SAE05] and Sadr, Anpalagan, and Raahemifar [SAR09], is the following: Assume a system with a set of users $\mathcal{U}$ and a set of orthogonal subchannels $\mathcal{N}$. The total bandwidth is $B$, the transmit power $P_{\text{total}}$ and the noise power spectral density $N_0$. The channel gain for user $u$ on subchannel $n$ is denoted by $h_{u,n}$. The variable $p_{u,n}$, with $p_{u,n} \geq 0$, denotes the power allocated to user $u$ on subchannel $n$. The flag variable $c_{u,n}$, with $c_{u,n} \in \{0,1\}$, indicates whether subchannel $n$ is allocated to user $u$. Using Shannon's formulation of channel capacity [Sha49], the rate of user $u$ can then be calculated as

$$r_u = \frac{B}{|\mathcal{N}|} \sum_{n\in\mathcal{N}} c_{u,n} \log_2 \left(1 + \frac{p_{u,n} h_{u,n}^2}{N_0 \frac{B}{|\mathcal{N}|}}\right). \tag{3.5}$$

Based on that, the optimization problem is defined as follows:

$$\max_{p_{u,n}, c_{u,n}} \quad \sum_{u\in\mathcal{U}} \mathrm{U}_u(r_u) \tag{3.6a}$$

$$\text{s. t.} \quad \sum_{u\in\mathcal{U}} \sum_{n\in\mathcal{N}} p_{u,n} \leq P_{\text{total}} \tag{3.6b}$$

$$\sum_{u\in\mathcal{U}} c_{u,n} = 1 \qquad \forall n \in \mathcal{N} \tag{3.6c}$$

Here, equation (3.6b) limits the total transmit power, and equation (3.6c) ensures that each subcarrier is used by a single user. Fairness can either be incorporated in the utility functions, or be formulated as additional constraints. Besides enforcing a fixed minimum rate per user, fairness can also be formulated as relations of user rates. In [SAE05; SAR09], the authors introduce predefined rate proportions $\gamma_u$ for each user $u \in \mathcal{U}$, and add an additional constraint:

$$r_{u_1} : r_{u_2} : r_{u_3} : \ldots = \gamma_{u_1} : \gamma_{u_2} : \gamma_{u_3} : \ldots \tag{3.7}$$

Extensions of this problem, including the allocation of groups of subcarriers, limited modulation order, and channel estimation errors, have, e. g., been studied by Huang et al. [Hua+09]. Fixing the power allocation allows to simplify the optimization problem, as performed by Margolies et al. [Mar+16]. Assume that in problem (3.6), $p_{u,n} = \frac{P_{\text{total}}}{|\mathcal{N}|}$. The feasible rate for user $u$ on resource $n$ can then be pre-calculated and is here denoted as $r_{u,n}$. This results in the following problem formulation (from [Mar+16], notation adapted).

$$\max_{c_{u,n}} \quad \sum_{u\in\mathcal{U}} \mathrm{U}_u \left( \sum_{n\in\mathcal{N}} c_{u,n} r_{u,n} \right) \tag{3.8a}$$

$$\text{s. t.} \quad \sum_{u\in\mathcal{U}} c_{u,n} = 1 \qquad \forall n \in \mathcal{N} \tag{3.8b}$$

Although simplified, this problem becomes NP-hard when binary RA is enforced (i. e. $c_{u,n} \in \{0, 1\}$) [Mar+16].

The previous problem formulations do not have a sense of time, i. e. they optimize RA for a single point in time only and derive rates from that. Different approaches can be used to incorporate time in the optimization. For a finite time horizon, resources can be enumerated and an all-encompassing problem can be formulated. In the special case of fixed power allocation, resources can be interpreted as bandwidth or time resources without influence on the problem formulation (see problem (3.8) and Margolies et al. [Mar+16]). Gradient-based approaches, in contrast, assume that infinite number of scheduling decisions are taken sequentially. When each step optimizes the projection of the users' rates onto the gradient of the sum utility, the system asymptotically approaches the optimum [Sto05].

When evolving time is part of the problem formulation, dynamics of user traffic become relevant. In contrast to the previous references, which adopted a full-buffer traffic model, optimization approaches can also be used for finite traffic. Jiang, Ge, and Li [JGL05] study the optimization of a queuing system with user-dependent channel quality. The authors derive RA strategies and weights from analytical solutions of the optimization problem. Proebster, Kaschub, and Valentin [PKV11] formulate an optimization problem to schedule the transfer of finite-size objects to users with time-varying channels. They also maximize sum utility, however their utility is defined as a function of the time the system takes to transfer an object. The completion times of transmissions are incorporated into the problem via finish time flags, which make the problem hard to solve.

## 3.4 Heuristics Used for Resource Allocation

Although RA can be formulated as optimization problem, solving such problems is typically not suitable for implementation in RT systems. Instead, heuristics are developed which allow computation of RA with reduced effort. Analytical evaluations and simulation studies are then used to assess performance and other properties of these heuristics. LTE implements RA centrally in the eNodeB. Therefore, distributed algorithms, like *Carrier Sense Multiple Access / Collision Avoidance* (CSMA/CA) used in *Wireless Local Area Network* (WLAN), are not discussed here.

### 3.4.1 Fixed Networks

Scheduling disciplines for fixed networks are a prominent research topic. An extensive overview over such disciplines, including classification, performance analysis, and discussion of implementation issues, is provided by Zhang [Zha95]. In principle, these approaches can also be used for wireless networks. However, the associated performance evaluations assume time-invariant and error-free transmission media, and are therefore not transferable to wireless networks.

Cao and Li [CL01] and Fattah and Leung [FL02] review approaches to adapt fixed network scheduling disciplines to wireless networks. They assume a two-state channel model, which either allows transmission (*good* state) or drops most or all packets (*bad* state). The general idea is to postpone users experiencing the bad channel state. Instead, the resources are assigned to other users, so that the system maintains high efficiency. Whenever a user's channel changes to the good state, he is compensated for the previous discrimination. Thereby, the approaches try to approximate the performance achieved in wired networks.

### 3.4.2 Maximizing Cell Metrics in Wireless Networks

RA heuristics have been designed for wireless networks in general and for LTE specifically. Capozzi et al. [Cap+13] give an overview over RA heuristics for LTE, also covering some scheduling disciplines for fixed networks.[6] According to the authors, most heuristics are based on the maximization of a per-PRB metric. This means that resource (or PRB) $n$ is allocated to user $u_n^\star$ by evaluating

$$u_n^\star = \arg\max_u \; m_{u,n}. \tag{3.9}$$

Here, $m_{u,n}$ is an abstract metric for user $u$ and resource $n$. Some heuristics extend this by an outer loop, which pre-selects users each TTI or on a coarser time frame. An alternative approach to the per-PRB selection is to sort users by an abstract priority, and then assign the preferred resources to them in this order. The further discussion focuses on heuristics based on a per-PRB metric following the notation of Capozzi et al. [Cap+13].

The simplest channel aware RA scheme is known as *Max C/I* or *Maximum Throughput* scheduler. The term *Max C/I* means maximizing the carrier-to-interference ratio. Thus, this heuristic assigns each PRB to the UE with the best channel conditions. Following the previous definitions, the per-PRB metric is defined by

$$m_{u,n}^{\text{Max C/I}} = r_{u,n}(t). \tag{3.10}$$

Without considering implementation restrictions (such as same modulation and coding for all PRBs allocated a UE), this maximizes the cell throughput. However, all resources are allocated to a single or a few users with the best channel conditions, while the other users get no resources assigned. The Max C/I scheme is therefore regarded as unfair.

A fairer RA heuristic for LTE networks is known as PF. Although in general the implementations used in commercial products are not publicly known, according to Margolies et al. [Mar+16], PF is the most prominent scheduler for LTE networks. It was patented by Chaponniere et al. [Cha+02]. The first documentation and evaluation in academic publications appeared in [JPP00]. The per-PRB metric is defined by

$$m_{u,n}^{\text{PF}} = \frac{r_{u,n}(t)}{\bar{R}_u(t-1)}, \tag{3.11}$$

where $\bar{R}_u(t-1)$ is the allocated rate filtered by a low-pass filter. It is calculated by

$$\bar{R}_u(t) = \left(1 - \frac{1}{t_c}\right) \bar{R}_u(t-1) + \left(\frac{1}{t_c}\right) r_u^{\text{act}}(t). \tag{3.12}$$

The value $r_u^{\text{act}}(t)$ is the rate actually allocated to user $u$ at time slot $t$. The parameter $t_c$ resembles the time constant of the low-pass filter. Stolyar [Sto05] proved that RA according to this algorithm asymptotically approaches the PF optimum defined by Kelly [Kel97].

[6] Other overviews over heuristics for opportunistic scheduling in OFDMA systems have been published by So-In, Jain, and Tamimi [SJT09] and Asadi and Mancuso [AM13].

Note that the original PF algorithm is defined for *time division multiple access* (TDMA) systems, i. e. without the index $n$. Thus, the filtered allocated rate is originally updated after each allocation. There are different options to transfer this to an OFDMA system such as LTE. $\bar{R}_u(t)$ can either be updated once per TTI as shown above, or repeatedly after each PRB has been allocated.

Additional parameters can be introduced to generalize the PF allocation heuristic. Wengerter, Ohlhorst, and Elbwart [WOE05] define the allocation metric as

$$m_{u,n}^{\mathrm{GPF}} = \frac{\left[r_{u,n}(t)\right]^a}{\left[\bar{R}_u(t-1)\right]^b}. \tag{3.13}$$

Setting $a = b = 1$ results in the original PF metric $m_{u,n}^{\mathrm{PF}}$. Setting $b = 0$ results in the Max C/I metric $m_{u,n}^{\mathrm{Max\,C/I}}$. Configuring $a = 0$ ignores the current channel situation. This scheme is known as *Blind Equal Throughput* (BET) and results in the same throughput transmitted to each user [Kel+08]. Further variants of the PF scheduler can be found in [Mar+16] and references therein.

### 3.4.3 Maximizing Individual QoS Metrics in Wireless Networks

The previously discussed RA heuristics opportunistically exploit the channel variations. However, they make no effort to deliver certain minimum performance to individual users. In LTE, some services require guaranteed rate or bounded packet delay (see section 3.2.1). QoS-agnostic RA heuristics are not suitable to satisfy such services [Cap+13]. Special algorithms have therefore been designed, over which this section gives an overview. The following paragraphs review heuristics which are based on a per-PRB metric as those discussed before. After that, approaches based on prioritization are presented.

Andrews et al. [And+01] propose a channel and QoS aware RA heuristic, which they term *Modified Largest Weighted Delay First* (M-LWDF). Although originally designed for a TDMA system, it can easily be adapted to the OFDMA nature of LTE. Following the previous notation, each PRB is assigned to the UE which maximizes the metric

$$m_{k,n}^{\text{M-LWDF}} = \alpha_k D_k^{\mathrm{HOL}} \frac{r_{k,n}(t)}{\bar{r}_k(t)}. \tag{3.14}$$

Here, $D_k^{\mathrm{HOL}}$ represents delay of the *head of line* (HOL) packet, i. e. the oldest packet in the queue. The variable $\bar{r}_k(t)$ resembles the short-term average of the possible data rate.[7] The factor $\alpha_k$ is a weight calculated by

$$\alpha_k = -\frac{\log \delta_k}{\tau_k}, \tag{3.15}$$

where $\tau_k$ is the delay threshold of UE $k$. The value of $\delta_k$ describes the acceptable probability that a packed exceeds the delay threshold and is dropped. A similar approach, based on a token counter, can be used to guarantee a minimum data rate instead of a maximum packet delay

[7] Note that, in contrast to the PF metric, the term $\frac{r_{k,n}(t)}{\bar{r}_k(t)}$ used here is independent of the previous allocations. Some authors, such as Capozzi et al. [Cap+13], use the PF metric instead. However, that does not reflect the original introduction by Andrews et al. [And+01].

[And+01]. According to Andrews et al. [And+01], M-LWDF can deliver all packets within the delay threshold if that is possible with any RA.

Another RA heuristic, giving increased priority to users with critical packet delays, is *EXP/PF* proposed by Rhee, Holtzman, and Kim [RHK03]. This heuristic divides users into RT and non-RT users. Two different metrics are calculated based on this differentiation:

$$m_{k,n}^{\text{EXP/PF}} = \begin{cases} \exp\left(\frac{a_k D_k^{\text{HOL}}(t) - \chi(t)}{1+\sqrt{\chi(t)}}\right) \frac{r_{k,n}(t)}{\bar{r}_k(t)} & \text{if } k \text{ is a RT user} \\ \frac{w(t)}{M(t)} \frac{r_{k,n}(t)}{\bar{r}_k(t)} & \text{otherwise} \end{cases} \tag{3.16}$$

Here, $a_k = \frac{c}{\tau_k}$, $c$ is a parameter, and $\chi(t)$ is calculated by

$$\chi(t) = \frac{1}{N_{\text{RT}}} \sum_{k \text{ is RT}} a_k D_k^{\text{HOL}}(t). \tag{3.17}$$

$N_{\text{RT}}$ represents the number of RT users. The fraction $\frac{r_{k,n}(t)}{\bar{r}_k(t)}$, which is used in the metric calculation for both types of users, resembles the current channel quality relative to the average channel quality.[8] For non-RT users, this is weighted with the factor $\frac{w(t)}{M(t)}$ This determines the relative weight of all non-RT users compared to the RT users. $M(t)$ is the average number of packets queued per RT user. A control loop adapts $w(t)$, so that $\chi(t) \approx c$:

$$w(t) = \begin{cases} w(t-1) - \epsilon & \text{if } \chi(t) > c \\ w(t-1) + \frac{\epsilon}{\kappa} & \text{otherwise} \end{cases} \tag{3.18}$$

The parameters $\epsilon$ and $\kappa$ (with $\kappa \gg 1$) control the convergence behavior. The authors also specify an alternative control loop, which adapts the weighting such that $\max_{k \text{ is RT}} D_k^{\text{HOL}} \approx \max_{k \text{ is RT}} \tau_k$. This RA heuristic allows RT and non-RT users to share the channel equally as long as the delay requirements are met. In case the delay becomes critical for a single user, the metric for that user is strongly increased.

Sadiq, Madan, and Sampath [SMS09] propose a similar approach, which they call *EXP rule*. They define the per-PRB metric as

$$m_{k,n}^{\text{EXPrule}} = b_k \exp\left(\frac{a_k q_k}{c + \left(\frac{1}{N}\sum_j a_j q_j\right)^{\eta}}\right) \cdot r_{k,n}, \tag{3.19}$$

where $b_i$, $a_i$, $c$, and $\eta$ are constant parameters. The value $q_k$ represents the queue length of user $k$. The authors state that the queue length can also be replaced with the delay of the HOL packet $D_k^{\text{HOL}}$. They propose to handle non-RT traffic in a similar way, but use a token counter instead of the real queues for those users. In their publication, Sadiq, Madan, and Sampath [SMS09] also propose another, logarithm-based heuristic. In addition, they evaluate the asymptotic behavior for increasing queue lengths.

[8] Rhee, Holtzman, and Kim [RHK03] as well as Capozzi et al. [Cap+13] state that this is equivalent to the PF metric, although the value is independent of previous allocations.

The previous references proposed the maximization of a per-PRB metric over all users, which is optionally calculated by a different formula for users with special QoS requirements. However, when many users are connected to a system, this can be computationally complex. The effort of calculating many per-PRB weights seems to be unnecessarily expensive especially in the case where, finally, all resources are assigned to few users with high priority.

To reduce this complexity, multiple two-step approaches have been proposed in literature [Mon+08; Zak+11]. The main idea is to have an outer scheduler, which pre-filters users once per TTI or on a coarser interval. Only users with high priority, either due to QoS requirements or channel conditions, are then considered in an inner scheduler. As the outer scheduler does not operate per PRB, the effort is reduced significantly.

One of these approaches is presented by Monghal et al. [Mon+08]. Their outer scheduler differentiates between users which received less than a target data rate and other users. The former have higher priority than the latter. Inside the sets, users are sorted by different metrics. Users below the target data rate are sorted by the BET metric, other users by a wideband version of the PF metric. The outer scheduler then selects a fixed number of users according to this prioritization. The inner scheduler assigns PRBs to these users. The authors evaluate different assignment strategies. These are based on PF or related per-PRB metrics. An additional weighting factor directs the inner scheduler to assign more PRBs to users which received less than the target data rate.

Zaki et al. [Zak+11] present a similar approach. They also group users into those with are to receive a guaranteed data rate and other users. However, their inner scheduler does not operate per PRB, but instead iterates over the list of pre-filtered users sorted by priority. To each user it assigns, from the set of free PRBs, that one with the best channel conditions. Users of the low priority group only get PRBs assigned after all users from the high priority group are satisfied. By this strict prioritization, it is easier to meet guaranteed rates.

## 3.5 Discussion

In the RA for wireless networks, partially contradicting objectives have to be balanced. Studies formulating RA as optimization problem apply different objectives. High-performance and reliable RA is important for vendors to differentiate from competitors. In addition, the computational complexity is also relevant, especially as systems have to scale to support hundreds and thousands of users. The RA heuristics become more complex by the introduction of carrier aggregation and more diverse capabilities of UEs.

RA has an influence on various components of a cellular radio system. Also, vice-versa, changing other system components may also change the optimal RA. Therefore, evaluations of cellular radio systems should always consider this mutual influence. Extending existing RA mechanisms by new aspects and components further increases their complexity. It can therefore be assumed that such extensions will only be adopted into products if the expected benefits are significant.

# 4 Efficient Operation in Multi-Cell Environments

In a cellular network, independent BSs serve disjoint sets of UEs. However, multiple BSs and UEs have to work together to provide services to the customers. The most prominent example for this is the *handover* procedure. When a user moves from the coverage area of one cell to that of another cell, the network hands over active sessions to provide continuous service without interruption. However, collaboration is also beneficial on the PHY layer. Inter-cell interference seriously impacts network performance. This effect can be mitigated by cooperation of neighboring BSs or special operations in the UEs. Furthermore, simultaneously serving a UE from multiple sites can improve reliability and efficiency. Mechanisms on PHY and MAC layer are complemented by methods for autonomous configuration and optimization.[1] Besides others, these mechanisms coordinate configuration of parameters for handover and interference mitigation. This thesis focuses on inter-cell interference and mechanisms for collaboration on the PHY layer.

Inter-cell interference occurs when the same resources are used by neighboring cells. It has a large impact on the SINR and thereby on channel capacity. Especially for users located close to the cell border, it is the major limiting factor for network performance. Depending on the desired fairness in the network, interference may also reduce the performance for other users. This is caused by the RA mechanism, which might allocate additional resources to cell border users to compensate for high interference. In contrast to intra-cell interference, which can be avoided by OFDM, in LTE inter-cell interference is not avoided by the design of the modulation scheme.

According to Cadambe and Jafar [CJ08], there are three ways to handle interference: It can be treated as noise, decoded, or orthogonalized. All three can be implemented at the receiver. In contrast, the transmitter can only orthogonalize its transmissions, or assist the receiver in one of the other ways.[2]

In LTE, resource usage in UL and DL is controlled by the eNodeB. Therefore, coping with interference at the transmitter often involves cooperation of BSs. This can be interpreted as weakening the cellular concept. Cooperation can either be statically configured or be automatically performed by the involved cells on different time scales. Besides reducing the impact of interference, cooperating eNodeBs can also jointly serve UEs. This approach is

[1] In LTE, this is often termed *self-organizing networks* (SON).

[2] This classification does not cover approaches such as dirty paper coding [Cos83; CS03]. Dirty paper coding requires up-front knowledge of the interfering signals at the receiver. It is not further discussed in this thesis.

known as CoMP.[3] UEs do not explicitly cooperate in LTE, as that would require out-of-band communication between them.

The following sections are structured as follows. Section 4.1 discusses mechanisms operating at the receiver side to cope with interference. Subsequently, section 4.2 introduces approaches to reduce interference by cooperation of transmitters. Finally, section 4.3 gives an overview over CoMP mechanisms, where cells cooperate tightly to achieve even higher network performance.

## 4.1 Coping with Interference at the Receiver

The simplest approach to cope with interference at the receiver is to treat it as noise. If the received interference power is weak compared to the received signal power, the impact on the performance is low. Yet, if the interference power is stronger, other approaches are more efficient.

Alternative approaches according to the classification by Cadambe and Jafar [CJ08] are the decoding of interference and the orthogonalization of desired signal and interference. These are both termed *interference cancellation*. Andrews [And05] gives an overview over that topic.

In contrast to treating interference as noise, decoding it makes use of the fact that it contains structure. The straight-forward approach is to jointly decode the desired and the interfering signals [Ver84; Ver86]. As this is computationally complex, approximations have been proposed, e. g., by Lupas and Verdu [LV89]. One of these approximations is to decode an interfering signal first and subtract it from the received waveform [PH94]. This is the same concept as the one introduced as SIC for MIMO operation in section 2.3.4.3. It is beneficial if the interfering signal is stronger than the desired signal.

Other approaches are based on orthogonalization. A receiver can, e. g., cancel out interfering signals in time domain based on their channel impulse response. Alternatively, the spatial domain (i. e. multiple receiver antennas) can be used to differentiate between the transmitter of the desired signal and interferers. In principle, a multi-antenna device receiving desired and interfering signals can be considered as spatial multiplexing system. Thus, approaches presented in section 2.3.4.3 can also be applied here.

Interference cancellation is efficient especially in the case where a single or a few strong interferers cause major impact. It can be performed locally and requires no coordination. It has therefore no impact on the standardization of the LTE air interface. However, its applicability in DL direction is limited by the high complexity, which increases device cost and power consumption of UEs.

In addition to purely local interference cancellation, LTE UEs can get assistance by their eNodeB. An eNodeB can provide information about reference signals transmitted by neighbor cells.[4] To ease the spatial filtering, an approach termed *interference alignment* strives for orthogonality of the desired and the interfering signals [MMK08; CJ08; APH13].

[3] In the extreme case, no independent cells would be visible to the UEs, or each UE would see its own cell. Such extreme cooperation is not covered by this thesis.

[4] This is introduced by Release 11 [3GPP 36.300 v11.4.0]. Assistance for interference cancellation is also covered by Release 12 under the term *network assisted interference cancellation and suppression* (NAICS).

## 4.2 Reducing the Impact of Interference by Coordinating Transmitters

Transmitters can combat interference by orthogonalizing their resource usage. In LTE this influences the RA for UL and DL transmissions, and therefore requires coordination of the involved eNodeBs. The main concept for orthogonalization is *spatial reuse*, which is introduced in section 4.2.1. The required coordination is discussed in section 4.2.2. The allocation of orthogonal resources to UEs in multiple cells can also be formulated as optimization problem, which is treated in section 4.2.3. Finally, section 4.2.4 gives an overview over standardization aspects for interference management in LTE.

### 4.2.1 Spatial Reuse of Resources

Spatial reuse describes the concept that closely spaced cells use orthogonal resources, while distant cells are allowed to reuse the same resources. In LTE, the two-dimensional resource grid (time and frequency, see section 2.4.1.2) can be used to orthogonalize transmissions. These resources are partitioned and neighboring cells use different partitions. As signals attenuate with distance, distant cells do not impact each other. The number of resource partitions determines the distance of cells using the same resources. This number is called *reuse factor*. In principle, both time and frequency domain can be used to orthogonalize resource usage. Traditionally, the frequency domain is used because that does not require synchronization of the BSs.

Classical networks like GSM employ *hard reuse* in the frequency domain. The available bandwidth is partitioned statically. Partitions are assigned to cells such that closely positioned cells use different resources. However, non-colliding assignments of reuse partitions to BSs is only possible for some regular layouts or high reuse factors. Also, it is often not possible that all BSs which use the same part of the spectrum have a uniform distance. This results in non-uniform interference. Although hard reuse reduces interference, it is inefficient, because each cell can use only a fraction of the available bandwidth. Therefore, one objective for the design of LTE was to be able to operate close to reuse factor 1, i. e. use all resources in all cells.

Multiple approaches were proposed to weaken the concept of hard reuse. A system can use lower reuse factors, which means that closer cells use same frequency band. However, that can lead to uneven distribution of interference, as some neighboring cells have to use the same resources. Instead of only using the allocated partition, a transmitter can also use the full bandwidth, but transmit with reduced power on all but the allocated partition. This concept is called *soft reuse*. The reduced transmit power causes limited interference, but allows to serve UEs with low channel attenuation. Another approach is to divide the available bandwidth into fractions, and then apply different reuse factors to the fractions. This approach is termed *fractional reuse*. UEs which suffer from interference can be served in a fraction with a high reuse factor. Other UEs are not impacted by the limited bandwidth.

In addition to orthogonalization on the spatial resolution of cells, it is also possible to steer the occurrence of interference with finer spatial granularity. In DL direction, multi-antenna arrays at the eNodeB can be used to steer beams away from UEs served by neighboring cells. In UL direction, the allocation of resources to UEs can be coordinated, such that UEs transmitting simultaneously can be distinguished by MIMO receivers of the respective eNodeBs. A related approach is interference alignment, which was introduced by section 4.1. In contrast to approaches operating on the time-frequency grid, approaches based on MIMO techniques require more detailed channel knowledge.

The global efficiency of spatial reuse depends on the relations between neighboring cells. When an interfering BS reduces transmit power on a resource, it suffers from reduced capacity. At the same time, UEs in multiple neighboring cells benefit. To make the reduction in transmit power efficient, these benefits have to compensate the reduced capacity at the interfering cell. Therefore, interference reduction can be especially efficient in heterogeneous cell layouts. There, a macro cell causes interference in many small cells. When the macro cell reduces its transmit power on a resource, all these small cells can use the respective resource more efficiently.

### 4.2.2 Coordination of Resource Usage

In case the resource usage is orthogonalized to reduce interference, the coordination of neighboring cells is beneficial. This is termed *interference coordination* (IfCo) or *inter-cell interference coordination* (ICIC). The schemes for coordination can be classified according to various aspects. The following paragraphs give an overview over these aspects.

The most prominent classification of IfCo approaches considers the time scale of the coordination. The simplest coordination is a static configuration. However, the optimal configuration of reuse factor, partitioning, power levels, etc. depends on the locations of the UEs and their data traffic. This varies over time and can only be estimated in advance. Therefore, a static configuration is sub-optimal in most of the time. In contrast to hard reuse, soft and fractional reuse allow an eNodeB to transmit on fractions of bandwidth with different characteristics. These schemes do therefore leave some degrees of freedom for the local RA mechanism. UEs can be moved between the fractions to adapt to changed requirements.

Dynamic approaches have the potential to adapt to instantaneous load situations. They can be further classified with respect to the time frame of adaptation. Slow adaptation eases the exchange of coordination messages. Faster adaptation allows to better adapt to instantaneous load and channel situations. Dynamic schemes require communication between eNodeBs.

IfCo schemes can also be classified regarding the amount and type of communication taking place. In the simplest systems, no communication takes place. While this still allows to achieve probabilistic gains from reduced interference,[5] it is typically not seen as IfCo scheme. Schemes based on implicit communication can be realized based on feedback from UEs, e. g. by using channel measurements or HARQ feedback to avoid resources with high interference [Kas+10]. Explicit communication allows to gain accurate information without much delay. However, communication introduces additional complexity into the system.

Another aspect suitable for classification is to consider what is influenced by neighboring cells. One group of approaches adapts only locally visible parameters (e. g. LA, allocation of resources to different UEs). Remotely visible decisions (e. g. transmit power, free resources) are taken solely based on local information. Thereby, these approaches do not reduce interference itself, but mitigate its negative impact on the network performance. However, they do not require

[5] For example, applying beamforming results in lower average interference for all users of a system.

joint decision making and do not balance conflicting objectives. They can thus be regarded as simple and robust.[6] Other approaches coordinate parameters that have influence on neighbors (e. g. transmit power, resource utilization). There, the potential for optimization is higher, but conflicting goals require a mechanism for joint decision making. As benefits and restrictions from IfCo can be unevenly distributed between cells, fairness can become relevant.

In case eNodeBs use explicit communication to coordinate parameters visible to each other, a mechanism for joint decision making is required. The simplest realization is a global coordinator. This receives information from all eNodeBs, derives the optimal IfCo parametrization, and communicates that to all participants. A global coordinator is difficult to implement in real systems. Therefore, a larger area can be split up into regions, and each region can be coordinated by a separate coordinator. This is sometimes termed decentralized coordination. Finally, distributed decision making is implemented by communication protocols between the eNodeBs, without support from additional devices [Nec09].

There are many examples in literature regarding heuristics for dynamic coordination. A well-structured overview over such approaches is provided by Necker [Nec09]. A framework and algorithms for the special case of heterogeneous networks are presented by Deb et al. [Deb+14].

### 4.2.3 Interference Coordination as Optimization Problem

Interference coordination with explicit communication and a central coordinator can be considered as optimization problem. This is an extension of the optimization problems introduced in section 3.3. Interference has to be added to the term that defines the rate achieved by a UE. Obviously, the problem has to include variables which influence interference. The discussion here focuses on power level IfCo, i. e. the variables that influence the interference define the transmit power of network nodes. However, also other variables can be used to influence interference, e. g. MIMO precoding matrices [Nec09].

A related problem is that where RA is not optimized, but each receiver is served by a different transmitter [SSM07; HBH06]. Also related is the extension to a problem including a variable association of UEs to cells [Deb+14; Mad+10].

Table 4.1 lists publications which formulate power level IfCo as optimization problem. They can be classified by how they model power allocation and the allocation of channel resources to participants of the network. They also differ by whether they assume constant or varying channel attenuation, and by their objective function.

The following paragraphs introduce an optimization problem which generalizes all the problems presented in these publications. The discussion focuses on the DL direction of a cellular network. It is assumed that there is a set of BSs $\mathcal{B}$ and a set of UEs $\mathcal{U}$. Each UE $u$ can be served by a single BS $b_u^\star$ only, but receives interference from all other BSs. The following hierarchical resource model is defined for the coordination of power levels and RA.

Transmit powers of the BSs are configured on the granularity of *power allocation resources*. Each BS uses a certain power level on a whole power allocation resource. If the power level

[6] One prominent example for this group is coordinated LA. Here, the selection of the MCS takes information about resource utilization at neighboring cells into account.

**Table 4.1:** Selected publications formulating power level IfCo as optimization problem

| reference | power allocation model | res. alloc. model | channel[a] | objective | notes |
|---|---|---|---|---|---|
| Das, Viswanathan, and Rittenhouse [DVR03] | single res., on-off power | selection of single UE | *n. a.* | weighted rates | BSs allocated to clusters, considers interferers in same cluster only |
| Li and Liu [LL03] | multiple res., fixed sizes, on-off power | selection of single UE | different | throughput | considers single strongest interferer only |
| Kim, Han, and Koo [KHK04] | multiple res., int.-sized, independent reuse factors | int.-sized | same | throughput, min. rates | precalculated rates per reuse factor |
| Huang, Berry, and Honig [HBH06] | single res., continuous power | fixed | *n. a.* | utility | no cells, but same resource used by all tx-rx pairs |
| Gesbert et al. [Ges+07] | single res., continuous power | selection of single UE | *n. a.* | throughput | |
| Assaad [Ass08] | two res., int.-sized, one with reuse factor | int.- or cont.-sized, see note | same | throughput, min. rates | each UEs assigned from one power res. only, selected by distance from BS |
| Rahman and Yanikomeroglu [RY10] | multiple res., fixed sizes, on-off power | selection of single UE | different | weighted rates | interferers restricted by distance and signal power,<br>but not incl. in problem formulation |
| Madan et al. [Mad+10] | multiple res., fixed sizes, continuous power | cont.-sized | same | utility | flexible cell assignment;<br>interferers restricted by signal power,<br>but not incl. in problem formulation |
| Deb et al. [Deb+14] | multiple res., int.-sized, limited on-off power | cont.-sized | same | weighted PF | heterogeneous scenario;<br>flexible assignment to macro/pico;<br>interference only from macro |

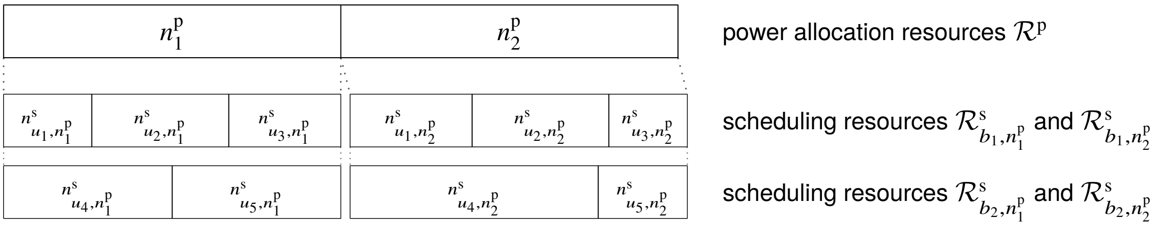

$n_1^\mathrm{p}$
$n_2^\mathrm{p}$
power allocation resources $\mathcal{R}^\mathrm{p}$
$n^\mathrm{s}_{u_1,n_1^\mathrm{p}}$
$n^\mathrm{s}_{u_2,n_1^\mathrm{p}}$
$n^\mathrm{s}_{u_3,n_1^\mathrm{p}}$
$n^\mathrm{s}_{u_1,n_2^\mathrm{p}}$
$n^\mathrm{s}_{u_2,n_2^\mathrm{p}}$
$n^\mathrm{s}_{u_3,n_2^\mathrm{p}}$
scheduling resources $\mathcal{R}^\mathrm{s}_{b_1,n_1^\mathrm{p}}$ and $\mathcal{R}^\mathrm{s}_{b_1,n_2^\mathrm{p}}$
$n^\mathrm{s}_{u_4,n_1^\mathrm{p}}$
$n^\mathrm{s}_{u_5,n_1^\mathrm{p}}$
$n^\mathrm{s}_{u_4,n_2^\mathrm{p}}$
$n^\mathrm{s}_{u_5,n_2^\mathrm{p}}$
scheduling resources $\mathcal{R}^\mathrm{s}_{b_2,n_1^\mathrm{p}}$ and $\mathcal{R}^\mathrm{s}_{b_2,n_2^\mathrm{p}}$

**Figure 4.1:** Hierarchical resource model for power level IfCo

is set to zero, it does not use the resource for transmitting. If it transmits, it can serve one or multiple receivers. Power allocation resources are orthogonal, i. e. they are separated by time, frequency, and / or code. They serve the coordination of resource usage and are therefore valid for all participants of the network.

Allocation of resources to receivers is modeled by *scheduling resources*. A scheduling resource defines a partition of a power allocation resource that is assigned to a single UE. In contrast to power allocation, RA is assumed to be invisible to other participants of the network. Thus, each BS independently splits each power allocation resource into multiple scheduling resources. Scheduling resources belonging to a single BS are orthogonal. It is assumed that a scheduling resource does not span multiple power allocation resources. Instead, a BS may serve the same UE on multiple scheduling resources.

In general, numbers and sizes of each type of resource are variable. However, a flexible number of resources with flexible characteristic is difficult to model. Therefore, all referenced sources use fixed numbers of resources and optimize their sizes and / or other properties (e. g. power levels, allocation to receivers).

The two resource types are here described as sets. The set of power allocation resources is denoted as $\mathcal{R}^\mathrm{p}$. For each power allocation resource $n^\mathrm{p} \in \mathcal{R}^\mathrm{p}$ and each UE $u$, there is one scheduling resource $n^\mathrm{s}_{u,n^\mathrm{p}}$.[7] The scheduling resources of a single BS are grouped as $\mathcal{R}^\mathrm{s}_{b,n^\mathrm{p}}$.

Figure 4.1 depicts an exemplary configuration of this resource model. There, two equally sized power allocation resources are defined. Therefore, each BS can use two different transmit powers. Two BSs $b_1$ and $b_2$ serve three and two UEs, respectively. They split the power allocation resources differently, i. e. their scheduling resources have different sizes.

The sizes of the resources (e. g. in units of time or frequency) are denoted as $s_{n^\mathrm{p}}$ and $s_{n^\mathrm{s}}$, respectively. The sizes of the scheduling resources have to sum up to the size of the respective power allocation resource.

$$s_{n^\mathrm{p}} = \sum_{n^\mathrm{s} \in \mathcal{R}^\mathrm{s}_{b,n^\mathrm{p}}} s_{n^\mathrm{s}} \quad \forall n^\mathrm{p} \in \mathcal{R}^\mathrm{p}, b \in \mathcal{B} \tag{4.1}$$

The channel attenuation is modeled by a factor for each combination of BS $b$ and UE $u$. This value may differ per power allocation resource $n^\mathrm{p}$ and is denoted as $h_{b,u,n^\mathrm{p}}$. The transmit power is

[7] An alternative formulation is also applied in literature. Scheduling resources are defined independently of UEs, e. g. with fixed size and count. A matrix of flag variables is then used to assign the scheduling resources to UEs. This is covered by the formulation here. See also equation (4.9).

configured separately per BS $b$ and power allocation resource $n^{\mathrm{p}}$. It is here denoted as $p_{b,n^{\mathrm{p}}}$. For each resource and BS, it must not exceed the power limit $p_{\max}$:

$$p_{b,n^{\mathrm{p}}} \leq p_{\max} \quad \forall n^{\mathrm{p}} \in \mathcal{R}^{\mathrm{p}}, b \in \mathcal{B} \tag{4.2}$$

Each UE can attain a different spectral efficiency on each power allocation resource. This depends on the channel gains and on the power allocations of the BSs. The derived data rate is the product of the spectral efficiency and the size of the allocated scheduling resource. It is calculated as

$$r_{u,n^{\mathrm{p}}} = r_{u,n^{\mathrm{s}}_{u,n^{\mathrm{p}}}} = s_{n^{\mathrm{s}}_{u,n^{\mathrm{p}}}} \operatorname{cap}\left( \frac{p_{b^\star_u,n^{\mathrm{p}}} h_{b^\star_u,u,n^{\mathrm{p}}}}{\sum_{b \in \mathcal{B}, b \neq b^\star_u} p_{b,n^{\mathrm{p}}} h_{b,u,n^{\mathrm{p}}} + N_0} \right). \tag{4.3}$$

The capacity function maps SINR to spectral efficiency. Most publications apply Shannon's [Sha49] channel capacity here (i. e., $\operatorname{cap}(\gamma) = \log_2(1+\gamma)$), but others leave the mapping undefined. The total data rate $r_u$ achieved by a receiver $u$ can be calculated by summing over all allocated scheduling resources:

$$r_u = \sum_{n^{\mathrm{p}} \in \mathcal{R}^{\mathrm{p}}} r_{u,n^{\mathrm{p}}} \tag{4.4}$$

In general, the system performance can be optimized by maximizing the utility $\mathrm{U}(\bullet)$, which is a user-specific function of the rate. Together with additional constraints, the utility function can serve to achieve a fair allocation of resources. Note that, in contrast to the fairness introduced in section 3.2.2, this defines fairness as a global objective. For example, if one BS serves many UEs, neighboring BSs may reduce interference to achieve balanced rates for all UEs served by all BSs. This results in the following optimization problem:

$$\begin{aligned} \max_{p_{b,n^{\mathrm{p}}}, s_{n^{\mathrm{p}}}, s_{n^{\mathrm{s}}}} \quad & \sum_{u \in \mathcal{U}} \mathrm{U}_u(r_u) \qquad (4.5) \\ \text{s. t.} \quad & \text{resource sizes, eq. (4.1)} \\ & \text{power limits, eq. (4.2)} \\ & \text{optionally fairness constraints} \end{aligned}$$

In literature, typically special cases of this generic problem are defined. A common simplification is to assume the same channel attenuation for all resources (denoted as *same* in column *channel model* of table 4.1). This can be reasoned, e. g., as coordinating only with the long term average channel state information.

The generic model allows to transmit with arbitrary power levels on different power allocation resources with variable sizes. This high degree of flexibility is not found in the models in literature. Some references restrict their models to a single power allocation resource [DVR03; HBH06; Ges+07]. Others restrict the sizes of the power allocation resources, either by fixing them [LL03; RY10; Mad+10], or by allowing only integer sizes [KHK04; Ass08; Deb+14].

A typical simplification is also to allow only predefined levels of transmit power, especially zero or full power (*on-off power* in table 4.1). Gesbert et al. [Ges+07] (see also [Gje+06]) have shown that this does not necessarily degrade the performance of the IfCo. This restriction makes it possible to enumerate the different levels of interference, and precalculate the users' rates. However, when there are many BSs in the system, the number of combinations can become prohibitive. Therefore, some of the referenced publications propose to restrict the set of considered interferers [DVR03; LL03; RY10; Mad+10; Deb+14].

Another prominent approach to further simplify the problem is to allow only some predefined patterns of active and inactive transmitters. These patterns can be defined by reuse factors as introduced in section 4.2.1. The assignment of reuse partitions to BSs can be treated as a separate optimization problem, which is not covered by the literature referenced here.

Another approach to simplify the problem is to select only a single UE per power allocation resource, and thereby omit the definition of scheduling resources. Typically, a matrix of flag variables for the allocation of power allocation resources to UEs is defined. For the generic model introduced here, this constraint can be formulated as

$$s_{n^{\mathrm{s}}} \in \{s_{n^{\mathrm{p}}}, 0\} \quad \forall n^{\mathrm{s}} \in \mathcal{R}^{\mathrm{s}}_{b,n^{\mathrm{p}}}, n^{\mathrm{p}} \in \mathcal{R}^{\mathrm{p}}, b \in \mathcal{B}. \tag{4.9}$$

This is even further restricted by Huang, Berry, and Honig [HBH06]. They do not model a cellular network, but a group of interfering transmitter-receiver pairs. Therefore, each transmitter serves a single receiver only, and there is no RA.

A related special case is when there are no fairness constraints, and the objective is to maximize the system throughput (see Gesbert et al. [Ges+07]). In that case, even if channel conditions between multiple resources differ, the optimal configuration for one resource is not influenced by other resources. Therefore, each power allocation resource can be optimized separately. Consequently, only one is included in the authors' problem formulation ($|\mathcal{R}^{\mathrm{p}}| = 1$). Similarly, a power allocation resource does not need to be split to be allocated to multiple receivers, but for each BS the UE with the best channel conditions is selected.

Although that is not directly visible from the problem formulations, time is modeled differently by the referenced literature. This influences how fairness can be achieved. When the long-term average RA is optimized, fairness has to be achieved during one execution of the optimization [Ass08; Mad+10; Deb+14]. However, some authors assume that the optimization is repeated for each time step [DVR03; RY10]. In that case, an outer control loop can, e. g., adjust weights which influence the optimization, and thereby achieve the desired fairness in the long run.

### 4.2.4 Standardization in LTE

LTE does not standardize an IfCo mechanism itself. However it includes some components which eNodeB vendors can use as foundation for their own implementations of IfCo. Some of these components target the air interface only, and assume that eNodeBs use vendor-specific mechanisms for coordination. Others cover signaling between eNodeBs, so that cooperating eNodeBs do not need to rely on non-standardized extensions. An overview over the standardized IfCo components in LTE given by [3GPP 36.300]. The following paragraphs summarize the main aspects targeted in different releases of LTE. Subsequently, mechanisms to provide flexibility to change the transmit power, mechanisms to measure interference channels, and signaling messages to be exchanged between eNodeBs are discussed.

#### *4.2.4.1 Concepts in Different LTE Releases*

In LTE Release 8, focus of standardization of IfCo components was to facilitate frequency domain IfCo. This imposes minimal requirements on the cooperation of the eNodeBs, as synchronization in time domain (i. e. alignment of the subframes) is not required.

Release 10 supports heterogeneous cell layouts, where multiple small cells are placed in the serving area of a macro cell. Here, special conditions render frequency domain IfCo unsuitable. First, there is a large difference in transmit power between macro cells and smaller cells. For UEs in the border area between two differently sized cells, there is a conflict regarding the optimal cell association: For DL transmissions, the received power should be maximized. In contrast, UL transmissions require low pathloss, because the transmit power of the UEs is limited. Second, system operators tend associate many UEs with small cells, to achieve balanced load although the sizes of the serving areas are different. Both results in strong interference. At the same time, IfCo in heterogeneous setups is often more efficient, because when a macro cell reduces transmit power multiple small cells benefit (see section 4.2.1). Thus, the Release 10 standardization enhanced the interference mitigation techniques and added time domain IfCo [Lop+11].

#### *4.2.4.2 Flexibility Regarding Transmit Power*

In the frequency domain, the transmit power can be coordinated either by statically configuring different carriers, or by reducing transmit power on some PRBs. The selective reduction of the transmit power can also be applied in the time domain. Multiple carriers provide no flexibility for coordination, as their configuration cannot be changed during operation of a cell. Therefore, the following discussion focuses on reducing the transmit power on a part of the resources of a single carrier.

To reduce the DL transmit power, reference signals, control channels, and data channels have to be regarded separately. The LTE standards prescribe constant transmit power for CRSs, so no IfCo is possible for those [3GPP 36.213]. For control and data channels, this depends on the used modulation scheme. For QPSK symbols, power can be adapted arbitrarily, as the receiver evaluates the phase of the received symbols, only. In contrast, for higher order modulation the receiver also interprets the amplitude by comparing it to the amplitude of the reference signals. When TMs relying on DM-RSs are used, the power of the reference signals can be adapted accordingly. However, when the receiver relies on CRSs, the difference in transmit powers of data and reference symbols has to be communicated explicitly. For most TMs, this is performed by higher layer configuration.[8] Thus, a different transmit power can be configured per UE, but that cannot be changed quickly. Compared to DL, more flexibility is provided for UL, because transmit power can be controlled separately for each UE. That comprises power for data and reference symbols.

For the PUSCH and the PDSCH, the RA mechanism can avoid resources with high interference. Special care has to be taken with control channels, which do not allow their location to be adapted flexibly. Therefore, resources where neighboring cells transmit with high power cannot

[8] A single exception is TM 5, which is intended for MU-MIMO operation and allows to specify an optional power offset for each RA.

be avoided. This is the main reason for the focus on time domain IfCo in LTE Release 10. There, *almost blank subframes* (ABSs) were introduced. These are subframes where control and data channels are transmitted either with reduced power or not at all. Thus, minimal interference is caused to neighboring cells. In DL, control channels also receive interference caused by full power CRSs from neighboring cells. This effect can be avoided by time-shifting the alignment of subframes in neighboring cells, such that control channels in one cell are not transmitted simultaneously with CRSs in another cell. Alternatively, interference cancellation can be used at the receiver to mitigate the impact of the reference signals.[9]

#### 4.2.4.3 Measurement of Interference Channels

In Release 8, normal mechanisms for CSI acquisition (see section 2.3.6) are also used to measure interference. As IfCo is assumed to be performed in the frequency domain, subband measurements can be used to decide whether it is efficient to serve a UE in that part of the bandwidth where interference is reduced. In Release 10, CSI measurements were extended to support time domain IfCo. Two subframe sets can be specified by the eNodeB. Measurements and reporting are performed separately. Release 11 brings additional flexibility by introducing CSI processes. There, separate resources can be configured to measure received signal power and interference. Neighboring cells can transmit CSI-RSs on those resources or leave them free, so that UEs can measure the case of no interference.

#### 4.2.4.4 Communication Among eNodeBs

Communication between neighboring eNodeBs uses the `X2` interface. Messages exchanged over this interface are defined in [3GPP 36.423]. In addition, vendor-specific extensions can be used to implement IfCo.

Signaling for IfCo among eNodeBs is part of LTE since Release 8. There, it targets power level IfCo in the frequency domain. Three types of `X2` signaling for IfCo are standardized [3GPP 36.423]. For DL, with *relative narrowband transmit power* (RNTP) messages eNodeBs inform their neighbors about the intended transmit power. This is signaled as a bit map, where each bit indicates high or low transmit power on a PRB, relative to an also communicated threshold. For UL, two signals are exchanged. The *overload indication* (OI) encodes the average received interference in three levels per PRB. The *high interference indication* (HII) notifies of high interference sensitivity. Neighboring cells are thereby requested to avoid allocation of selected resources to cell border UEs.

Since Release 10, standardization includes mechanisms to support time domain IfCo. There, eNodeBs exchange patterns of normal subframes and ABSs [3GPP 36.423 v10.1.0]. To allow stable configuration of CSI measurement, a subset of the ABSs is defined as measurement subframes. While a subset of ABSs may change to adapt to load variations, those denoted as measurement subframes are intended to be more permanent. To support joint adaptation of the ABS patterns, feedback messages allow to communicate ABSs utilization and efficiency.

[9] See introduction of interference cancellation in section 4.1.

In Releases 12 and 13, extended signaling for time and frequency domain IfCo was standardized under the term *Inter-eNodeB CoMP*[10] [3GPP 36.300 v12.8.0]. Neighboring eNodeBs exchange hypotheses and associated benefits. Here, a hypothesis comprises two-dimensional patterns of power-reduced PRBs for one or multiple cells. The exchanged messages can be used by a distributed decision making process to coordinate RAs.

## 4.3 Coordinated Multipoint Transmission and Reception

The approaches discussed in section 4.2 focus on the reduction of interference between neighboring cells. There, each cell is equivalent to a sector of a site and serves a distinct set of UEs. However, also a tighter cooperation is possible, where single UEs are served simultaneously from multiple sites. In the LTE context, this termed CoMP. Note that approaches to reduce interference, which are based on cooperation of neighboring cells, can also be classified as CoMP. However, these are omitted here because they were covered by section 4.2.

### 4.3.1 CoMP Cell Configurations

In CoMP, a sector at a site is not necessarily equivalent to a cell. Therefore, in accordance with the 3GPP documents [3GPP 36.819 v11.2.0], the following discussion uses the term *point* to describe an array of antennas mounted at the same site and with the same directional characteristic. For the special case of DL transmissions, a point is also termed *transmission point* (TP). A point can correspond to a sector of a macro site, which is mounted above rooftop and transmitting with high power. However, a point can also be transmitting with much lower power and be mounted below rooftop or indoors.

In a traditional (non-CoMP) setup, a UE receives control channels and data from the same TP, which corresponds to its serving cell. That point also receives data in UL direction. This interrelation between the TP used for control channels and the TP used for data channels is removed in a CoMP setup. The control channels corresponding to a cell can be transmitted from one or multiple TPs. At the same time, a TP can transmit or receive data belonging to UEs associated with different cells. In a simple CoMP setup, the control channels and most of the data channels of a cell are transmitted by a single TP. Service for some UEs in the cell border area is supported by a neighboring TP, which also serves its own cell. In a more advanced setup, a single cell employs multiple TPs to serve a large area. MU-MIMO and dedicated control channels (EPDCCH) are used to maximize network performance.

### 4.3.2 CoMP Approaches for Data Transmission and Reception

Lee et al. [Lee+12], and similarly also the 3GPP [3GPP 36.819 v11.2.0], classify CoMP into three types of approaches. The first is *coordinated scheduling* (also *coordinated beamforming*). This was already covered by section 4.2. The remaining two approaches are *transmission*

[10] In 3GPP terminology, there is no clear distinction between IfCo and CoMP. See also section 4.3.

*point selection* and *joint transmission and reception*. These two are discussed in the following paragraphs.

In dynamic point selection, a UE is served by a single point at a time, however that can change from subframe to subframe without handover. This applies for DL as well as for UL transmissions. Dynamic point selection exploits channel variations experienced by UEs in cell border areas. To decide which point to use, the quality of both channels has to be estimated. For advanced MIMO schemes, additional information is required for both channels.

A tighter cooperation of the points is assumed for joint reception and joint transmission. Joint reception means that for decoding data transmitted in UL direction, the signal received at multiple points is combined. This is transparent to the transmitting UE. Analogously, for joint transmission multiple TPs transmit simultaneously on the same resources to a single UE. This can be performed either coherently or non-coherently. Coherent operation aims at constructive superposition of the received signals at the UE. This results in good performance, however requires detailed channel information and exact synchronization of the TPs to align the phases of the signals. In contrast, non-coherent operation is easier to realize. It can still utilize power and diversity gains.

Note that CoMP and IfCo approaches do not exclude each other, but can be combined to improve efficiency. For example, two points can simultaneously serve a single UE by using joint transmission, while a third point mutes the respective PRBs to avoid to cause interference.

The performance of CoMP setups has been studied for multiple years. Overviews are given by Lee et al. [Lee+12], Maattanen et al. [Maa+12], and Marsch and Fettweis [MF11]. In addition, the CoMP study item in 3GPP Release 11 provides definitions and performance evaluations [3GPP 36.819 v11.2.0].

### 4.3.3 Standardization in LTE

CoMP is introduced in LTE Release 11 with a focus on the air interface. It is assumed that cooperating points are handled by the same eNodeB (e. g. sectors of the same site), or that vendor-specific extensions are applied. Although Releases 12 and 13 introduce signaling for CoMP on the X2 interface, that targets interference reduction, only. This was covered by section 4.2.4.

UL CoMP is mostly transparent to the involved UEs. The required CSI can be acquired by measuring the same SRSs at multiple points. This can be used, together with a prediction of the received interference, to derive the LA parameters. In case multiple UEs transmit simultaneously to multiple points, the eNodeBs can configure different UE-specific sequences of DM-RSs to better separate the received signals.

For DL CoMP, UEs are configured to use TM 10. That allows to define multiple CSI processes. Each process can measure CSI-RSs transmitted by different antenna ports. These ports can be located at different sites. This allows to derive all required MIMO parameters, e. g. the optimal precoding matrix for a beam transmitted jointly from two TPs. As TM 10 relies on DM-RSs for MIMO decoding and demodulation, the applied CoMP scheme does not need to be communicated to the UE explicitly. To realize CoMP gains also for DL control channels, the EPDCCH can be used, which also employs on a form of DM-RSs.

## 4.4 Summary and Discussion

This chapter gave an overview over different approaches to make efficient use of the radio channel in a multi-cell scenario. Receiver side actions are used more heavily in UL direction, because they cause significant computational effort. Their application is mostly transparent to the network. The capabilities of the receiver can, however, be utilized by the transmitter. The BS can possibly optimize IfCo by allowing not none, but a single strong interferer. Similarly, interference alignment is based on the fact that receivers can cancel interference.

Orthogonalization of interfering signals is a reliable approach to improve cell-edge performance. To avoid disadvantages for cell center users, fractional reuse is typically applied. Traditional deployments are based on static or semi-static configuration of reuse patterns. Dynamic coordination in distributed deployments of eNodeBs requires communication via the X2 interface. This often implicates a significant delay. Therefore, fast coordination, e. g. as part of RA in 1 ms interval, is difficult to realize. This is, however, different in C-RAN deployments. There, RA for multiple cells is performed in the same cluster of BBUs, which allows for low-latency communication. In addition, in case the centralized BBUs are built by the same vendor, coordination does not need to rely on standardized interfaces.

Similar to IfCo, CoMP can improve the throughput for cell edge users. It requires tight cooperation of the involved BSs. This is, again, difficult to realize in distributed deployments.

IfCo and CoMP have different effects on the compute effort in a BBU pool. First, for both there is a certain effort for the coordination of power levels, RA, and MIMO parameters. However, the effort for encoding and decoding can be assumed to be higher with CoMP, because more antennas are involved and optionally non-linear algorithms are executed (e. g. dirty-paper coding, SIC). In contrast, with IfCo encoding and decoding have a similar effort as without IfCo. Only increased data rates for cell edge users can result in higher effort. In the special case where IfCo disables the transmission on some resources (on-off power), the compute effort for encoding and decoding is reduced.

# 5 Elastic Utilization of Compute Resources

Chapter 2 introduced LTE and the standardized transmission technology. Subsequently, chapters 3 and 4 provided an overview over approaches for RA and efficient operation in multi-cell environments, respectively. Based on these, this chapter presents the proposed mechanism to elastically adapt the compute resource utilization of a mobile communication system.

First, section 5.1 motivates the chosen approach, derives research questions, and highlights the contributions of this thesis. Subsequently, this work is categorized in the context of related work from different research areas in section 5.2. The system model for the following evaluations is then specified in section 5.3.

The design of the proposed system is based on a sound analysis of the interrelations between RA, IfCo, LA, computational effort, and network performance. These interrelations are studied by solving different variants of an optimization problem. This problem is defined in section 5.4, and its solutions are discussed in section 5.5. The proposed system is then derived from the findings in section 5.6. While a general discussion of the properties of the proposed system is included therein, an in-depth performance analysis is documented in the subsequent chapter 6.

A preliminary version of the optimization problem defined in section 5.4 and its evaluation (similar to section 5.5) has been published in [Wer15].

## 5.1 Problem Statement

Based on the background in technology and research presented in the previous chapters, this section defines the actual problem statement of this thesis. It is split into two parts. First, section 5.1.1 motivates that a mechanism for elastic utilization of compute resources is required to design efficient C-RAN systems. Subsequently, section 5.1.2 details the research questions and contributions that were introduced in section 1.4.

### 5.1.1 Motivation

For a long time, the compute effort caused in wireless communication devices did not concern. Instead, the efficient utilization of the available spectrum was the predominant objective for the design of wireless networks [Vri+02; 3GPP 36.913]. However, this thinking begins to change.

Installation and operation of cellular networks also contributes to the total costs of network operators. While efficient spectrum utilization is still an important goal, new aspects came up in the last years. These comprise energy efficiency [Pic+08; Cor+10; HBB11; EARTH], flexible programmability [SN06; I+14b; Zhu+11], and efficient operation in C-RANs [I+14a; CMRI11; Wüb+14]. These new aspects are related to the estimation and limitation of the compute effort caused in the RAN. The following paragraphs focus on the flexible implementation on GPPs and on the operation as virtual functions hosted by a cloud of hardware resources.

When baseband processing uses ASICs, a significant share of the total cost is fixed (e. g. engineering, production of photolithography masks), and installing more compute power comes at a relatively low additional cost [KMN02]. Also, ASICs typically have a good relation of compute power per consumed energy, so that additional compute power does not significantly increase total energy consumption. To some extent, this still applies to more flexible hardware such as FPGAs and DSPs. However, it is not true for freely programmable GPPs. While there the fixed costs are lower, compute capacity can be enlarged only by buying larger or more compute chips. This significantly contributes to total cost and power consumption [SN06]. Therefore, the proper dimensioning of compute capacity is crucial for GPP implementations.

For the comparison of ASIC and GPP implementations, Li et al. [Li+11] differentiate two aspects of complexity. ASICs and other specialized hardware can efficiently handle computation complexity. This means that they are good at delivering numerical performance. In contrast, GPPs are superior in coping with structural complexity. Implementing multiple and structurally complex algorithms in software causes less effort than realizing the same in hardware. In addition, flexible software implementations cause no or only minimal drawbacks at runtime, while flexible hardware implementations require more chip area and thereby contribute to the unit costs. This benefit of GPPs is utilized in this thesis. The proposed system comes with additional structural complexity, but allows an eNodeB implementation to cope with reduced computational capabilities.

The required compute power for baseband processing in an eNodeB depends on multiple varying parameters and is therefore difficult to predict [WGP13]. The installed compute capacity either needs to be dimensioned for peak load, or a potential shortage of compute capacity has to be accepted. Dimensioning for peak load can be inefficient and overly expensive [Ros+15a]. This is in particular true for large pools of baseband units, where the peak load is much higher than the typical load [Bha+12; WGP13; RTV15]. Elastic utilization of compute resources allows to cope with resource shortages [Ros+15a]. Therefore, compute resources can be dimensioned tighter, which saves hardware and energy.

One of the main desires which cause the trend to C-RANs is to operate a RAN with high efficiency and flexibility. This is equivalent to the cloud idea in IT installations [CMRI11]. There, *virtual machines* (VMs) replace hardware servers as structuring elements of the IT architecture.[1] A large number of VMs is aggregated on a cluster of hardware units. The hardware can be operated with a high utilization, because the aggregation results in a more constant load. In addition, the load can be balanced between hardware units by live migration of VMs. This allocation flexibility also allows to clear hardware units, which can be used for maintenance or to switch them off to

[1] The term VM is here used as synonym for virtualization on different layers, encompassing system and container based virtualization [Sol+07].

save energy. In case of hardware failures, the same flexibility supports to quickly restore services on unaffected hardware units.

In IT clouds, load and performance are monitored and actions (e. g. moving VMs, booting additional hardware) are taken reactively. This is possible because IT applications are typically elastic, i. e. they can tolerate moderate overload. In case of a shortage of compute capacity, requests take longer to complete, but the service is not disrupted immediately. To apply the same concept to RAN operation, the latter has to support elastic utilization of compute resources.

For the implementation of an eNodeB, strict timing requirements apply [I+14a; PHT16; Ros+15a]. To support these on GPP platforms, special actions have to be taken. For example, Tan et al. [Tan+11] assign dedicated processor cores to signal processing tasks. While typical IT operating systems and virtualization platforms are not designed to give RT guarantees, there are some specialized products and approaches which do so [Gua+16; GCL14]. However, these cannot manage overload. In case of overload, it is not possible to prolong the response times of the system. Instead, the RAN software has to explicitly manage overload to guarantee in time transmission of standardized signals. Doing so efficiently is the major objective of this thesis.

The LTE standard requires an eNodeB to transmit reference and synchronization signals at fixed intervals. In addition, it requires to react on messages within a certain time (e. g. send a HARQ acknowledgment). Further, for high performance network operation, various other tasks such as RA and encoding of user data have to be performed. To conform with the standard also in case of overload, an eNodeB is required to transmit all mandatory signals and messages. However, that alone is not sufficient to maintain high system performance. Instead, most of the non-mandatory data transmissions should be continued, so that the system makes the most of the available compute resources.

This can be illustrated by the following examples. A simple overload management can be realized by stopping all non-required interactions. An eNodeB can fall back to transmit only reference and synchronization signals and stop allocating new resources to UEs. In addition, processing of ongoing transmissions can be aborted, because all participants of an LTE network can cope with temporal outage. Such handling of overload implicates a significant impact on network performance. A more elaborate overload handling mechanism can, e. g., allocate only a part of the time-frequency resources to new transmissions. In addition, it can tune LA parameters so that baseband processing becomes computationally more efficient.

The efficiency of the overload handling determines the optimal dimensioning of the hardware resources. If the system cannot cope efficiently with overload, such situations have to be avoided, and a significant safety margin has to be applied. When the system operates efficiently in overload situations, this safety margin can be reduced. If moderate overload causes no significant impact on system performance, a C-RAN can be operated in permanent overload. The system can thereby achieve optimal utilization of the compute hardware.

### 5.1.2 Research Question and Contributions

The main research question of this thesis is how to achieve an efficient and elastic utilization of compute resources. When overload can be handled efficiently, good system performance is maintained under limited compute resources. Section 1.4 introduced three objectives. First, the potential efficiency of an LTE network coping with compute resource overload is to be assessed. Second, a simple overload handling mechanism is to be designed, which achieves good efficiency and is suitable for implementation in a real LTE system. Third, the efficiency of the designed mechanism has to be shown.

The first objective for this thesis is to evaluate the potential efficiency of coping with compute resource overload. Thereto, an optimization problem is formulated and solved, which covers the most important variables and their influence on system performance and compute effort. Namely, the problem encompasses IfCo, RA, and the configuration of LA parameters. To be solvable, this optimization problem necessarily has to be abstract and simplify some effects. Nevertheless, the evaluation of solutions to this optimization problem allows to estimate the compute capacity required to deliver a certain system performance. This estimation is independent of potential restrictions incurred by implementations of overload handling mechanisms. In addition, it allows to gain insight into the influence of different variables and allows to infer guidelines for the design of the overload handling mechanism.

The second objective is to design an overload handling mechanism, which mediates between network performance and implementation complexity. In addition, one of the design goals is to avoid overly radical changes to already existing building blocks of the C-RAN system. The design of the proposed heuristic is well-founded on insights inferred from solutions to the optimization problem. The heuristic is based on an efficiency metric which assesses the effect of LA decisions on compute requirements and network performance.

Finally, the proposed overload handling mechanism is evaluated thoroughly. It is thereto compared to a simple baseline heuristic and to the optimal solution from an adapted optimization problem in a simplified scenario. In addition, it is evaluated in a more complex, dynamic scenario. While an optimal solution cannot be used as a benchmark there, the second scenario serves to verify that the heuristic can cope with dynamic effects occurring in a real system.

An all-encompassing evaluation of the components of a mobile communication system exceeds the scope of this thesis. Thus, some restrictions have to be defined.

In general, similar effects influence compute effort in UL and DL processing. In both directions, RA and LA parameters define the resulting processing effort. Also, for both directions, time-frequency resources without transmission are beneficial for the interference situation. Therefore, leaving resources free to simultaneously reduce interference and compute requirements is a promising approach. The effort for UL processing is expected to be larger than that for DL processing [I+14a; Des+12]. The efficiency of overload handling in UL direction has already been studied by Rost, Talarico, and Valenti [RTV15].[2] Therefore, the model and all evaluations in this thesis focus on the compute effort required for signal processing in DL operation. It is, however, assumed that the same general approach can also be applied to design an overload handling mechanism for UL processing.

The evaluated DL processing blocks encompass FEC processing, modulation, and MIMO operations. The effort for the inverse DFT, which is also part of signal processing for OFDM transmissions, is not considered. This is reasoned by the fact that DFT calculation imposes a

[2] See also [Ros+15b] and the discussion of related work in section 5.2.3.

constant effort. On one hand, it cannot be influenced by changing RA or LA parameters, and therefore it is difficult to adapt it to limited compute capacity. On the other hand, this effort can be easily predicted, so no safety margins are required for hardware dimensioning. It is assumed that higher layers (i. e. processing of the PDCP) and control plane operations cause only small effort compared to the main data traffic transmitted via the PDSCH.

This thesis evaluates the efficient handling of overload using the example of signal processing for LTE DL processing. The same approach can also be applied to evolved 5G networks, given that suitable compute effort models become available.

## 5.2 Classification of Related Work

This thesis proposes a mechanism to achieve elastic utilization of compute resources in a mobile communication system. In doing so, it focuses on the compute load caused by signal processing. The related work in the area of efficient and adaptive signal processing is presented in section 5.2.1. The underlying compute system can be considered as a RT system. In that area of research, adaptive complexity is a known method to cope with overload. The related approaches are discussed in section 5.2.2. Achieving elasticity eases the management of the processing resources and allows for a tighter dimensioning of these resources. An overview over approaches for the management and dimensioning of compute resources in C-RAN systems is given by section 5.2.3. Finally, the discussion of related work is summarized in section 5.2.4.

### 5.2.1 Efficient Signal Processing in Communication Systems

Efficient signal processing is a major objective for the design of wireless communication systems. Overly complex signal processing requires powerful hardware, which increases system cost and energy consumption. For example, one of the reasons to apply OFDM modulation in LTE is that this eliminates the need for complex time domain equalization.

Efficient signal processing is a sub-aspect of the energy efficiency of cellular networks. The energy efficiency has to be considered separately for BSs and UEs. On the BS side, especially for macro BSs, the energy consumed by power amplifiers dominates the other components of the system [Cor+10; HBB11]. However, signal processing contributes significantly to the power consumption in the UEs. This is caused by the fact that a UE transmits only rarely and on a fraction of the bandwidth, while it listens to the radio channel more often.[3] Besides cellular networks, energy efficiency is also a prominent topic in sensor networks [CGB04; CGB05; Jay04]. The following discussion focuses on efficient signal processing.

The computational complexity in an LTE system can be reduced directly on the level of the signal processing algorithms. However, it is also influenced by mechanisms and configuration on higher layers. These approaches are detailed in the following two sections 5.2.1.1 and 5.2.1.2, respectively. The selection of algorithms and the configuration of parameters is typically static.

[3] A UE has to decode the PDCCH to detect when it has to receive data. If that is the case, it also decodes the respective parts of the PDSCH. In addition, it is required to measure the radio channel. UEs can switch to power saving modes to reduce this effort.

In contrast, this thesis targets the dynamic adaptation of the system to the available compute resources. Section 5.2.1.3 covers related work which also considers dynamic reconfiguration.

#### *5.2.1.1 Algorithms with Reduced or Configurable Complexity*

The design as well as the implementation of signal processing algorithms allow to trade communication performance off for reduced computational complexity. Li et al. [Li+11] provide a structured overview over these approaches.

A prominent example for a set of algorithms performing the same task with different performance and complexity are MIMO decoders (see also section 2.3.4.3). Simple linear decoders have lowest complexity, but also limited performance. The ML decoder is known to be optimal but involves a search over all combinations of possibly transmitted symbols, which prohibits its application in many use cases. Advanced MIMO detectors such as sphere decoding can achieve performance close to ML with significantly reduced computational complexity [Zim07; CTL12].

There are also signal processing algorithms where the complexity can be influenced by configuration parameters. For example, Desset et al. [Des+11] present a configurable MIMO decoder. There, a configurable search range allows to reduce complexity at the cost of lowered detection performance. Similarly, the number of iterations performed by a turbo decoder influences detection performance and computational complexity.

Besides the design of the algorithms, a similar trade-off can be made during implementation of the algorithms. An algorithm can be implemented with fixed point or floating point arithmetic. In addition, the accuracy of the representation of the numbers can be chosen. Fixed point arithmetic with low accuracy is simpler to realize in hardware. However, it is more susceptible to numerical errors and therefore reduces the performance of the communication system [Jan+11].

Approaches on signal processing level are not evaluated in this thesis. However, the general trade-offs realized by configurable signal processing are similar to those studied in this thesis. Thus, the approach of this thesis, which consists of selecting configurations by individual computational efficiency is potentially also applicable for these trade-offs.

#### *5.2.1.2 Mechanisms and Configuration on Higher Layers*

An LTE system comprises of many mechanisms on different layers. Some of these can be adapted or parametrized to allow to reduce signal processing effort.

When a UE has not transmitted or received data for some time, it enters the *discontinuous reception* (DRX) mode. In this mode, it decodes the PDCCH only in predefined subframes. Thus, DRX allows to switch off the receiver in the remaining subframes. The time after which DRX mode is entered and the subframes where the UEs decode the PDCCH can be configured by the BS. DRX can significantly reduce the signal processing effort at the UEs. However, it introduces additional delay for the transmission of data in DL direction.

Another approach to reduce computational effort of a receiver is to limit the number of iterations of the turbo decoder. When a MCS is selected for an allocated channel resource, this determines the amount of encoded data. The closer this is to the capacity of the resource, the more iterations are required to successfully decode the data at the receiver. Thus, maximizing the total throughput of a system results in complex decoding. Valenti, Talarico, and Rost [VTR14] propose to take decoding complexity into account for the MCS selection. By selecting more robust MCSs, their system reduces the required number of iterations.

The BS can select the MIMO modes used in UL and DL direction. This is performed individually for each UE according to its velocity and the characteristics of the radio channel (see section 2.3.4.4). However, MIMO modes also come with different measurement and signaling overhead and with different computational complexity at the UE and at the BS. They can thus be used to trade throughput off for reduced signaling overhead or for reduced complexity. For example, Cardoso et al. [Car+13] propose to change the number of active transmit antennas at the BS. They show that using simpler MIMO modes reduces power consumption when the network load is low.

RA also has a significant influence on the signal processing effort. Signal processing only occurs for those resources which are used for transmission and reception. The benefit of unused resources is also considered by Cardoso et al. [Car+13]. When resources are allocated, the associated processing effort often depends on the radio channel and thereby on the served UE. This is similar to the proposal of Kim et al. [Kim+09], where a resource is allocated based on a metric which includes the energy efficiency of the transmission. Summarizing, leaving resources free or serving UEs with low processing requirements can serve to reduce the total computational effort. However, this impacts the network throughput and fairness.

In this thesis, the influence of two higher level mechanisms on the computational complexity is evaluated. These are the selection of MIMO modes and the RA. In addition, the effect of reduced interference is studied, which can be seen as additional benefit of not allocating resources.

#### *5.2.1.3 Dynamically Trading Network Performance Off for Reduced Complexity*

The previous sections described different components that can be adjusted to reduce computational complexity. Such adjustment is typically performed during the design of the system. It is, however, also possible to change the configuration at runtime to adapt to varying requirements or conditions. Approaches which do this are discussed in the following paragraphs.

Desset et al. [Des+11] propose to scale the accuracy of their configurable MIMO detector to not exceed the allowed complexity. The same MIMO detector is evaluated jointly with a flexible turbo decoder by Desset and Torrea Duran [DT12]. They study the power consumption of a UE in different scenarios. The authors show that, while the influence of the flexible turbo decoder is limited, the adaptive MIMO decoder allows to save energy by efficiently utilizing the radio channel. However, they do not describe a mechanism which dynamically adapts the decoder.

Kim et al. [Kim+09] propose an approach for energy-efficient resource allocation and MIMO mode selection for UL transmissions. They assume that multiple UEs dynamically start new transmissions in a cell. For each UE, their algorithm strives to realize the required rate with the lowest energy consumption. When the load in the cell is low, simpler transmission schemes can be used to save energy.

Dynamically choosing how to balance between computational complexity and network throughput is also the objective of a publication by Li et al. [Li+11]. Their system is based on signal processing blocks which allow to scale signal processing quality and computational complexity. During operation of the system, a controller adapts the configuration of these blocks to achieve a certain desired quality. Their publication contains three case studies, where this general scheme is applied for different components of a signal processing system.

The discussed publications propose different approaches how to dynamically trade off network throughput for computational complexity or power consumption. They do all focus on a single communication link with predefined requirements. In contrast, this thesis simultaneously decides this trade-off for multiple competing links. This can be considered as RA problem, similar to those discussed in chapter 3 for allocation of channel resources [Ros+15a]. Further approaches in this direction are discussed in section 5.2.3.

### 5.2.2 Real-Time Scheduling

The communication protocols impose strict timing requirements on the computation for cellular communication systems. Each BS has to transmit its radio frames at fixed intervals. The specifications of the radio interface do not allow to delay their transmission, e. g. to cope with computational load peaks. The implementation of a BS can therefore be considered as RT system. The following section 5.2.2.1 introduces RT systems. Subsequently, section 5.2.2.2 discusses approaches to cope with overload in such systems.

#### *5.2.2.1 Introduction to Real-Time Systems*

According to Sha et al. [Sha+04], "a RT system is one with explicit deterministic or probabilistic timing requirements." It handles multiple jobs competing for one or multiple shared resources, e. g. a processor. Each job is characterized by the point in time when it enters the system (its *release time*), the point in time when it has to be completed (its *deadline*), and a duration of resource utilization (its *execution time*) [But+10]. The point in time when a job is completed is known as *finish time*. A sequence of related jobs is termed task, process, or thread.

In a mobile communication system, the signal processing for each UE in each subframe can be considered as separate job. The execution time of each job is variable, because it depends on the number of allocated PRBs and other parameters. Collating multiple jobs to tasks is not useful here, because jobs do not have any common characteristics.

It is assumed that all jobs associated with the same subframe have the same release time and the same deadline. Further, it is assumed that this deadline occurs before the release time of the subsequent subframe. Thus, at no time jobs of consecutive subframes are simultaneously schedulable. This means that processing resources cannot be shifted between subframes, and each subframe can be analyzed individually.

Buttazzo et al. [But+10] classify the timing requirements of tasks as *hard*, *firm*, *soft*, and *non-RT*. A hard task requires that each job is completed within its deadline. A firm task allows that a limited fraction of the jobs miss their deadline. The performance of a soft task degrades gracefully when the completion is delayed. Finally, non-RT tasks state no constraints on their finish times. A single system can, in principle, serve tasks with different requirements. However, that is not further considered here.

The signal processing of a mobile communication system can be interpreted as having hard or firm requirements. The classical approach is to assume hard timing requirements, i. e. designing the system such that all signal processing can be completed in time. However, in this thesis, the system is assumed to have firm timing requirements. This means that it is accepted that the signal processing for some UEs is not completed within the deadline. The completion time of a job is only relevant to determine whether the respective deadline is met. This imposes a significant simplification compared to typical RT scheduling problems, because it is not relevant in which order the jobs are processed.

#### *5.2.2.2 Coping with Overload*

Hard RT systems guarantee to meet all deadlines. Such a guarantee can be provided by a static *feasibility analysis* (also termed *schedulability analysis*). Thereto, worst-case conditions (e. g. regarding release and execution times) are assumed for all tasks. An overview over different tests, which can then be applied to the set of tasks, is given by Sha et al. [Sha+04]. Even for simple algorithms, complex computer systems with hierarchical memory architectures result in variable execution times. Consequently, it is difficult to determine the worst case execution time of a job. Thus, guaranteeing hard RT requirements often results in inefficient utilization of the compute resources [But+10].

In contrast to hard RT systems, firm and soft RT systems can cope with overload by skipping or delaying tasks, respectively. The performance of such systems can be defined by assigning a *utility* to each job, which is a function of its finish time [But+10]. For firm tasks, this has a constant value when the job is completed before the deadline, and drops to zero afterwards. For soft tasks, the utility decreases monotonically when the completion is delayed. The performance of the system is then the sum of the utilities of all jobs.

Firm and soft RT systems can cope with jobs for which the worst case requirements cannot be stated in advance. To maximize system performance, an algorithm dynamically decides the acceptance and the schedule of jobs during system operation. In case the arriving jobs can all be served within their deadline, the algorithm defines the order of their serving. Otherwise, the system has to cope with an overload situation. As the order of completion is not relevant here, the following paragraphs focus on the handling of overload.

Sha et al. [Sha+04] as well as Buttazzo et al. [But+10] state three approaches to reduce load in an overload situation: Skipping jobs, increasing their *interarrival times* (IATs), and reducing their execution times. Skipping is applicable to firm jobs and implies a degradation of the total utility. The remaining two approaches rely on an adaptive job model.

The main approach of the system proposed in this thesis is to reduce the execution times of jobs. Whenever that is not sufficient to prevent overload, the system skips jobs as additional measure. The utility function, which is used as objective function in this thesis, is a function of transmitted data rates. It is not related to the utility as a function of job finish time in RT systems.

The adaptation of job execution times is typically modeled as follows [Shi+89; Liu+91; Liu+94; Sha+04; But+10]. Each job is divided into a mandatory and an optional sub-job with individual execution times and the same deadline. All mandatory sub-jobs have to meet their deadlines. An optional sub-job becomes executable only after the respective mandatory sub-job has been completed. The optional sub-job can be skipped, executed partially, or executed completely. The performance of the system is measured in terms of a total error. The error of a job depends on the fraction of the optional sub-task that is completed in time. This approach is known as the *imprecise computation* model.

The imprecise computation model allows to interrupt the optional sub-job at any time. It is thus applicable whenever an algorithm successively improves the precision of the result. For the signal processing in mobile communication systems, the parameters affecting the execution time of a job are decided before the execution starts. Thus, this model is not applicable here.

An alternative model for adaptive execution times is the one introduced as *multiple methods* by Garvey, Humphrey, and Lesser [GHL93]. There, multiple methods or algorithms can be used for the same job, each making a different trade-off between solution quality and time. The authors focus on a sequence of dependent jobs that has to be completed in order. They propose a scheduling algorithm which is based on pruning dominated methods.

The multiple methods model is applicable for the signal processing of a mobile communication system as evaluated in this thesis. When a job comprises the encoding of data for a single UE, different methods can be, e. g., using different MIMO modes. These result in different execution times and data capacities. The jobs used to encode data for different UEs are independent, which simplifies their scheduling. In contrast to the publication by Garvey, Humphrey, and Lesser [GHL93], this thesis also considers the fairness of the resource allocation. Further, the objective is here to design a distributed system, while Garvey, Humphrey, and Lesser [GHL93] propose a centralized approach similar to the optimization problem defined in section 5.6.2.

The distributed system proposed in this thesis is based on a prediction mechanism for the efficiency threshold. This can be interpreted as a simple control loop which modifies the load of the system. This is similar to a method known as *feedback scheduling*, which has been proposed by Stankovic et al. [Sta+99].[4] There, a RT scheduler is enhanced by a control loop. The authors provide an example, where the control loop switches between multiple methods and adapts an admission control mechanism to achieve a deadline miss ratio close to zero.

While their objective is the same as that used in this thesis, their system model differs. They model the set of queued jobs as a liquid tank, where new jobs are added and completed jobs are removed continuously. This differs significantly from the model used here, where new jobs are computed at each subframe and no jobs can persist in the system. While their system has an internal state (the queued jobs), the system proposed here is intrinsically stateless.[5] The only state used here is the artificially added threshold variable, which is adapted at every subframe. In addition to these differences, Stankovic et al. [Sta+99] do not detail how to select the jobs which are switched to computationally simpler methods. Performing this selection efficiently in a distributed manner is a central component of this thesis.

[4] See also references therein as well as in [Årz+00] and [But+10, chapter 8].

[5] More accurately, a new set of jobs is created for each subframe. The sets of consecutive subframes do not intersect. However, the data traffic and the radio channel impose a correlation on the execution times of the jobs of consecutive subframes.

### 5.2.3 Dimensioning and Management of Compute Resources in C-RANs

When a C-RAN is implemented on GPPs, the signal processing has to be handled by a RT computer system. Therefore, both areas of research have to be combined when considering the dimensioning and management of compute resources in C-RANs.

In a classical mobile communication system, various trade-offs between network performance and computational complexity are decided when the system is designed. The computational resources are then dimensioned for peak load [Wüb+14]. This approach is well suited for systems where the variation of the computational load is low, e. g. because the system only uses two antennas. In addition, this approach is favorable for systems based on ASIC or FPGA implementations [KWM11]. There, computational power can be realized efficiently. On the contrary, structural complexity, e. g. due to support of adaptive mechanisms, can become a burden.

C-RAN systems can better adapt to variable computational requirements. Virtualization facilitates a flexible allocation of functions to hardware. Thereby the computational load of multiple hardware units can be balanced. In addition, with software implementation on GPPs, structural complexity is easier to handle. This allows, e. g., to implement multiple variants of algorithms and switch between these to balance between network performance and computational complexity. Thus, as described in section 5.2.1.3, the average computational load can be reduced by adapting the performance of the network to the requirements of the respective link.

The beneficial effects of a C-RAN system on the computational load are shown by Werthmann, Grob-Lipski, and Proebster [WGP13]. The publication shows how multiplexing of the computational load of multiple BSs in a single large BBU pool reduces the variance of the computational load. By serving a large area, such a BBU pool also balances the spatial inhomogeneity of the load. This allows to save processing hardware compared to a dimensioning for peak load. These evaluations are based on percentiles of the measured load. However, the publication does not cover a method to cope with compute resource overload.

Similarly, Wang et al. [Wan+16] propose a dynamic allocation of processing functions to hardware. Their objective is to power down unused hardware and thereby save energy. However, they do also not consider the case that the available hardware does not satisfy the current load.

Pompili, Hajisami, and Tran [PHT16] propose an approach to predict and monitor the required processing resources. Their system makes use of repeating patterns of user activity, e. g. by recording diurnal load. Based on that, it predicts the required processing resources. It can, however, not cope with overload situations. Instead, the authors state that so much hardware has to be provisioned that the processing of each frame is completed before deadline. This is difficult to realize in practice, because a tight upper bound for the future load cannot be given easily. Guaranteeing that the resources are always sufficient thus results in inefficient over-provisioning.

The concept of *computational outage* has been introduced by Valenti, Talarico, and Rost [VTR14] (see also [RTV15]). This describes the probability that the available compute resources are not sufficient to handle the instantaneous load. It thereby accepts that some blocks of scheduled radio resources or transmitted data have to be discarded due to shortage of compute capacity. To minimize the influence of this on the system, the authors propose to order the processing such that those UEs with high channel quality, which also cause high effort, are skipped first.

Another simple approach to handle overload is proposed by Werthmann et al. [Wer+15]. There, a central instance simultaneously reduces the number of PRBs allocated to each UE until the total load can be handled by the available resources.

While these approaches allow a system to cope with overload, it is more efficient to apply one of the adaptive mechanisms discussed in section 5.2.1. Such a system is described by Rost et al. [Ros+15b]. They propose two heuristics to select the MCSs used for UL transmissions. As explained in section 5.2.1.2, their system switches to more robust modes to reduce the computational complexity of the turbo decoder. It can thereby satisfy a constraint on the computational load. The proposed heuristics operate after time-frequency resources are allocated to UEs. Both are centralized, i. e. the transmissions of all UEs in the system are handled by a joint procedure. They do not consider fairness of the compute resource allocation, but strive to maximize the sum rate of the system.

The major objective of this thesis is to maximize the efficiency of a mobile communication system in cases of computational overload. Thereto, the configuration of the system is constantly adapted. This is similar to the proposal of Rost et al. [Ros+15b]. However, in contrast to that publication, this thesis aspires a fair allocation of compute resources and presents a simple distributed heuristic. While Rost et al. [Ros+15b] study the effort for decoding received data, here the effort for signal processing necessary to transmit data is evaluated.

Rost et al. [Ros+15a] state that the allocation of radio and compute resources should be managed jointly. This thesis follows that advice in the formulation of the optimization problem in section 5.4. However, the related evaluations in section 5.5 show that adapting MIMO modes alone achieves comparable performance. Therefore, the joint allocation of both types of resources is not performed by the proposed system.

### 5.2.4 Summary

The previous sections gave an overview over the related work from three different subjects. Section 5.2.1 discussed approaches for efficient and adaptive signal processing in mobile communication systems. There exist various proposals to reduce signal processing complexity, e. g. by using approximating algorithms, simplified implementation, or avoiding complex operations by modifying higher level mechanisms. However, all these impact the network performance. These trade-offs between complexity and performance are typically chosen at design time of the system. In contrast, they are here performed during system operation. The existing approaches for such dynamic selection were discussed in section 5.2.1.3. Unlike this thesis, however, they do not focus on a total compute resource limitation of a system, but on single links with predefined network performance requirements.

Section 5.2.2 treated the mobile communication system as a RT system. Such systems can have different timing requirements. The timing requirements of the mobile communication system are here assumed to be firm. That allows some tasks to miss their deadline. The order in which the jobs are served is not relevant here, because all jobs of a subframe have the same arrival time and deadline. However, known mechanisms for coping with overload can be applied. In the system proposed in this thesis, the processing of some jobs is skipped. In addition, an adaptive job model is used, where the execution times of the jobs can be reduced. Here, the multiple methods model applies, which allows to select one of multiple predefined methods (or implementations) to complete a job. In contrast to this thesis, the solution for this model proposed in literature describes a central solving algorithm and does not consider fairness.

Finally, section 5.2.3 targeted the management and dimensioning of compute resources in C-RANs. In classical mobile communication systems, compute resources are dimensioned statically to match the requirements in all situations. Different publications have shown the inefficiency of this approach. Some of the alternative approaches accept that compute resources are not sufficient in all situations. In such cases, some transmissions are skipped. However, only a single publication suggests to use the adaptivity of the system to make most efficient use of the available compute resources. While that publication targets UL operation, this thesis focuses on DL processing. There, an ad-hoc design of two centralized heuristics is presented. In contrast, a more fundamental approach is followed in this thesis.

Summarizing, all three subjects provide a basis for this thesis. However, dynamically adapting the trade-off between computational complexity and network performance in a distributed system, at the same time striving for fairness and maximizing performance, has not been performed before.

## 5.3 System Model

The system model applied here for optimization and simulation largely conforms to the 3GPP simulation guidelines in [3GPP 36.814]. An urban macro network is modeled, which is described as *Case 1* in that document. The scenario is simplified in some aspects to allow solving the optimization problems. The most relevant model parameters are listed in table 5.1. The following sections provide an in-detail description of the model components.

First, section 5.3.1 describes how geometry is modelled. Subsequently, section 5.3.2 defines two different cell layouts and introduces the modeling of BSs. The placements and antennas of UEs are defined in section 5.3.3. Sections 5.3.4 and 5.3.5 deal with the model for the radio channel and the applied abstractions for PHY layer mechanisms, respectively. The two data traffic models used are defined by section 5.3.6. The model for the processing effort is derived from literature in section 5.3.7. Finally, section 5.3.8 describes the applied simplifications regarding RA.

### 5.3.1 Geometry Model

The geometry model serves as base for storing and calculation of spatial positions, distances, and angles. All geometric operations are based on a three-dimensional Cartesian coordinate system. However, the use of the third dimension is limited. BSs and UEs are positioned with fixed heights in relation to ground. Here, ground is a flat surface which is equivalent to the x-y plane. This means that there are neither geographic nor architectural sources for elevation.

A simulation model can only encompass a scenario of limited size and complexity. In the straight-forward approach, BSs and UEs are placed in a bounded area. However, in that case border effects occur, because units close to the border of the area see, e. g., no interference originating from outside the area. To avoid these distortions, a wrap-around scenario is used

**Table 5.1:** System model parameters on the basis of 3GPP *Case 1* [3GPP 36.814]

| Property | Configuration |
|---|---|
| Geometry model | 3D coordinate system, wrap-around |
| Cell layout | 7 or 19 tri-sectorized sites, inter-site distance $d_{\text{is}} = 500\,\text{m}$ |
| BS / UE height | 32 m / 1.5 m |
| Carrier configuration | FDD, 10 MHz bandwidth per carrier, center frequency 2 GHz |
| Resource granularity | $N_{\text{PRB}} = 50$ PRBs per subframe |
| BS TX power | 46 dBm |
| Penetration loss | 20 dB |
| Path-loss | $128.1 + 37.6 \cdot \log_{10} d$, in dB; $d$: distance in km [3GPP 36.814] |
| Shadow fading | log-normal with 8 dB standard deviation |
| MIMO channel model | Spatial Channel Model [3GPP 25.996], UE velocity 3 km/h |
| BS antenna elements | 3D radiation pattern with 15° tilt [3GPP 36.814] |
| UE antenna elements | omnidirectional |
| BS antenna array layout | 8 elements in 4 cross-polarized pairs with $0.5\,\lambda$ distance |
| UE antenna array layout | 4 elements in 2 cross-polarized pairs with $0.5\,\lambda$ distance |
| UE antenna noise figure | $F = 9\,\text{dB}$ |
| MIMO receiver | zero-forcing |

here. Wrap-around here means that the scenario is cloned and shifted in a repeating pattern. A hexagon shape of the scenario with six shifted clones is used. Everything which leaves the scenario enters a shifted clone. This is equivalent to re-entering the original scenario on the opposite edge. The same is applied to radio propagation.

The wrap-around principle is depicted in an example in figure 5.1. There, a UE moves out of the scenario area (orange) on the north-east side. It enters a shifted clone of the scenario (dashed orange), and thereby the scenario itself again at the south edge.

Two configurations of the scenario are defined, which correspond to two BS layouts with seven and 19 sites. For a detailed definition of the wrap-around geometry see appendix A.

### 5.3.2 Cell Layout and Base Station Model

In accordance with [3GPP 25.814] and [3GPP 36.814], BS sites are placed as follows. The first site is placed at the origin of the coordinate system, i. e. at $(0\,;0)$. Further six sites are placed around that in the first tier, such that the distance between sites is $d_{\text{is}}$ and one site is placed at $(0\,;d_{\text{is}})$. For the larger scenario, twelve sites are placed in the second tier, such that one is placed at $(0\,;2d_{\text{is}})$. For the smaller scenario, the second tier is omitted.

At each site, three sectors are served as independent cells. The set of all cells is denoted as $\mathcal{B}$. The sector antennas are mounted in a height of 32 m and their main lobes point to 30°, 150°, and 270°. Together with the wrap-around scenario defined in the previous section, this results in a regularly repeating pattern. Therefore, simulation results for all cells are equivalent except for

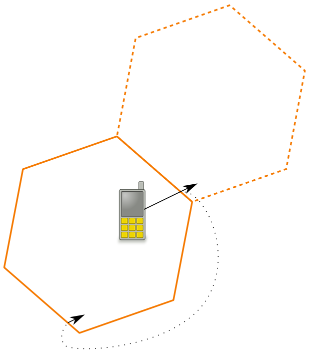


**Figure 5.1:** Example for wrap-around principle

statistical variations. Positions of sites and orientations of sectors are depicted in figures A.1 and A.2 (see page 184).

For the calculation of attenuation, the three dimensional antenna pattern defined in [3GPP 36.814] is applied. The tilt of the antenna is configured to 15°,[6] which is also specified there for calibration purposes. Note that smaller tilt angles increase interference, and render IfCo more advantageous.

For modeling of MIMO effects, the configuration of the antenna array of each sector has to be specified. This is not defined in the simulation scenario description in [3GPP 36.814]. However [3GPP 36.814, table A.3-1] gives multiple exemplary configurations, which are originally intended for IMT-Advanced evaluations. From those, configuration $E$ is applied here. This defines eight antennas in four groups. The antennas are placed on a horizontal mount which is orthogonal to the main lobe direction. The distance between the groups is $0.5\lambda$, with $\lambda \approx 0.15\,\mathrm{m}$ for 2 GHz carrier frequency. The two antennas of each group are polarized with 45° and −45°, respectively.[7] This layout is depicted in figure 5.2.

To make use of this antenna setup, LTE TM 9 is used (see section 2.3.4.4). Here, BSs use precoding matrices from the codebook specified by the LTE standard, because that is also used for CSI reporting.

To reduce compute effort, the concept of virtual antennas is introduced, which allows BSs to use less antennas for transmission. The BS can combine multiple physical antennas to a single virtual antenna by transmitting the same signal on them. Thereby, simpler precoding matrices can be used, which reduces the processing effort.[8] At the same time, the full transmit power of the system can be used, as all physical antennas are used for transmission. The modeled MIMO channel suffers from correlations, but the correlation between the two polarization planes is lower than that between arbitrary antennas. Therefore, it is most efficient to keep the polarization planes separated even when using four or two virtual transmit antennas. To transmit on four virtual antennas, each of these pairs transmits the same signals: ($A_1$, $A_2$), ($A_3$, $A_4$), ($A_5$, $A_6$), ($A_7$, $A_8$). Two virtual antennas are formed by the quadruplets ($A_1$, $A_2$, $A_3$, $A_4$) and ($A_5$, $A_6$, $A_7$, $A_8$).

Each cell transmits a total power of $P_{\text{total}} = 46\,\text{dBm}$, which is distributed equally to the antennas.

[6] This means 15° downwards relative to the horizontal orientation.

[7] According to Dahlman, Parkvall, and Sköld [DPS16], the LTE precoding codebook for eight antennas is made for "closely spaced cross-correlated" antennas.

[8] The benefit of this flexibility is highlighted by figure 5.14 in section 5.5.3.

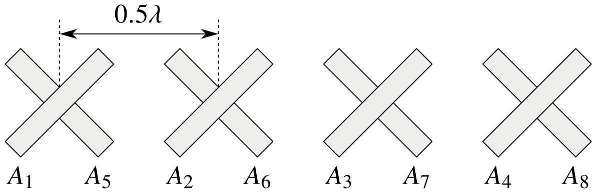


**Figure 5.2:** Layout of the BS antennas

### 5.3.3 User Model

The set of all UEs in the system is denoted as $\mathcal{U}$. The number of UEs in the scenario is either fixed or variable, which depends on the applied traffic model (see section 5.3.6). The UEs are placed randomly. No UEs are placed in circles of size 35 m around the BS sites. Except that restriction, coordinates are distributed uniformly over the whole area of the scenario, with a fixed height above ground of 1.5 m. Especially, there is no fixed number of UEs placed in the service area of each BS.

During evaluation, UEs do not move. However, depending on the traffic model, UEs may be created and removed dynamically. For modeling of the multipath propagation in the channel model, a velocity of 3 km/h with random direction is assumed.[9]

The UE antennas are omnidirectional and have a noise figure of $F = 9\,\text{dB}$. For MIMO effects, it is assumed that each UE has four antennas in two cross-polarized groups with $0.5\lambda$ distance between the groups. The orientation of the antenna array is selected randomly.

### 5.3.4 Channel Model

The channel model describes the propagation of radio waves between transmitters (BSs) and receivers (UEs). The same model is applied for all transmitters and receivers, i. e. the propagation of interference is modeled in the same way as the propagation of the desired signals.

The radio channel is modeled as combination of two groups of effects, the macro-scale attenuation effects and the small-scale fading effects. Here, the macro-scale attenuation effects are a scalar which depends on the relative positions of BS and UE. As UEs don't move during simulation, this value is constant for each pair of BS and UE. In contrast, the small-scale fading differs per antenna in a MIMO antenna array, and also per subcarrier. The small-scale fading is modeled as a time- and frequency-dependent complex matrix with zero mean and unit power. The two effects are combined by multiplication:

$$\mathbf{H}_{u,b}(t, f) = \sqrt{\gamma_{u,b}}\, \widehat{\mathbf{H}}_{u,b}(t, f) \tag{5.1}$$

Here, $\mathbf{H}_{u,b}(t, f)$ is the time-varying radio channel between BS $b \in \mathcal{B}$ and UE $u \in \mathcal{U}$ as function of time $t$ and frequency $f$. It is defined as the product of the constant macro-scale attenuation $\sqrt{\gamma_{u,b}}$ and the time-varying, frequency selective small-scale fading $\widehat{\mathbf{H}}_{u,b}(t, f) \in \mathbb{C}^{N_{\text{Rx}} \times N_{\text{Tx}}}$. Here,

[9] This approach, i. e. fixed locations, but velocity assumed for fast fading effects, is proposed for simulations with the *Spatial Channel Model* [3GPP 25.996, section 5.1].

$\gamma_{u,b} \in \mathbb{R}^+$ is the attenuation of the signal power. It consist of pathloss, penetration loss, shadow fading, and antenna patterns, i. e.

$$\gamma_{u,b} = \gamma_{u,b}^{\mathrm{PL}}\ \gamma^{\mathrm{P}} \gamma_{u,b}^{\mathrm{SF}}\ \gamma_{u,b}^{\mathrm{Ant}}. \tag{5.2}$$

The pathloss $\gamma_{u,b}^{\mathrm{PL}}$ is calculated as

$$10 \log_{10} \left( \gamma_{u,b}^{\mathrm{PL}} \right) = -1 \left( 128.1 + 37.6 \log_{10} \left( \frac{d_{u,b}}{1000} \right) \right), \tag{5.3}$$

where $d_{u,b}$ is the distance between $u$ and $b$ in meters [3GPP 36.814; 3GPP 25.814]. In addition, a constant penetration loss of 20 dB is assumed for each channel realization, i. e. $\gamma^{\mathrm{P}} = 0.01$.

The shadow fading $\gamma_{u,b}^{\mathrm{SF}}$ is modeled as log-normally distributed random variable with unit mean and a standard deviation of 8 dB. Two UEs at the same position experience the same shadow fading. This correlation decreases exponentially with distance. At a distance between the UEs of 50 m the coefficient of correlation has fallen to 0.5. From two sectors of the same site, a UE experiences the same shadow fading. The fading from sectors of different sites is correlated with a coefficient of correlation of 0.5.

The last component of the attenuation effects is the attenuation caused by the radiation patterns of the transmit antennas,[10] denoted as $\gamma_{u,b}^{\mathrm{Ant}}$. The model calculates the relative angle between the main lobe of the BS antenna and the direction in which the UE is seen from the BS. The gain of the antenna is then determined by the formulas given in [3GPP 36.814, table A.2.1.1-2].

The multipath propagation, the resulting fast fading, and its correlation between multiple antennas is modeled by the *Spatial Channel Model* (SCM) defined in [3GPP 25.996], here operated in the *urban macro* configuration. The calculated channels are normalized such that they have zero mean and unit power, because all effects except fast fading are already modeled by $\gamma_{u,b}$ as described before. For simplicity, we take only one sample of the multipath propagation model per PRB pair. The used implementation of the SCM is based on that of Salo et al. [Sal+05].

### 5.3.5 Model for Interference and Physical Layer Processing

Assume that each UE $u$ is served by a single BS $b_u^\star \in \mathcal{B}$. This is defined to be the one from which the UE receives the strongest signal, i. e. $b_u^\star = \arg\max_{b \in \mathcal{B}} \gamma_{u,b}$. The UE can receive interference from all other BSs, which are here denoted as $\mathcal{I}_u = \mathcal{B} \setminus b_u^\star$.

Based on this and on the introduction of MIMO in section 2.3.4, the vector of received symbols at UE $u$ in a precoding MIMO system can be formulated as

$$\mathbf{y}_u = \mathbf{H}_{u,b_u^\star} \mathbf{W}_{b_u^\star} \mathbf{x}_{b_u^\star} + \sum_{i \in \mathcal{I}_u} \mathbf{H}_{u,i} \mathbf{W}_i \mathbf{x}_i + \mathbf{n}. \tag{5.4}$$

[10] As described above, only the BSs use directional antennas, while UE antennas are modeled as omnidirectional.

Here, $\mathbf{W}_b$ denotes the precoding matrix used by BS $b$ and $\mathbf{x}_b$ the symbols transmitted by the same BS. Note that this applies for all subcarriers and OFDM symbols, however that has been omitted from the notation for simplicity.

We do here assume that perfect CSI is available at both transmitter and receiver. The ZF approach is used to reconstruct the transmitted signals. This multiplies the received symbols with the pseudo-inverse of the effective channel $\mathbf{G} = \left(\mathbf{H}_{u,b_u^\star}\mathbf{W}_{b_u^\star}\right)^\dagger$. The resulting estimate of the transmitted symbols is

$$\begin{aligned}
\tilde{\mathbf{x}}_{b_u^\star} &= \mathbf{G}\mathbf{y}_u \\
&= \mathbf{G}\left(\mathbf{H}_{u,b_u^\star}\mathbf{W}_{b_u^\star}\mathbf{x}_{b_u^\star} + \sum_{i\in\mathcal{I}_u}\mathbf{H}_{u,i}\mathbf{W}_i\mathbf{x}_i + \mathbf{n}\right) \\
&= \mathbf{G}\mathbf{H}_{u,b_u^\star}\mathbf{W}_{b_u^\star}\mathbf{x}_{b_u^\star} + \sum_{i\in\mathcal{I}_u}\mathbf{G}\mathbf{H}_{u,i}\mathbf{W}_i\mathbf{x}_i + \mathbf{G}\mathbf{n}
\end{aligned} \tag{5.5}$$

It can be assumed that the effective channel is invertible, because otherwise the transmitted signal cannot be reconstructed accurately and the transmitter should use a different precoding or number of spatial layers. Therefore, $\mathbf{G}\mathbf{H}_{u,b_u^\star}\mathbf{W}_{b_u^\star} = \mathbf{I}$, i. e. the effect of the channel can be reverted and the transmitted spatial layers can be separated. Equation (5.5) simplifies to

$$\tilde{\mathbf{x}}_{b_u^\star} = \mathbf{x}_{b_u^\star} + \sum_{i\in\mathcal{I}_u}\mathbf{G}\mathbf{H}_{u,i}\mathbf{W}_i\mathbf{x}_i + \mathbf{G}\mathbf{n}. \tag{5.6}$$

The objective of the PHY layer model is to determine the data capacity and decode probability for each set of allocated resources. Given the size of the allocated resources, the data capacity can directly be calculated from the selection of modulation scheme and code rate. The decode probability then also depends on the received signal quality, which is influenced by channel realization, received interference, and MIMO configuration.

Detailed modeling of the PHY layer procedures is computationally complex and therefore typically not performed in system level simulations. Nevertheless, realistic modeling of the spectral efficiency of different MIMO modes with the same channel is required here. Therefore, an abstraction proposed by Colom Ikuno [Col13] is applied, which is summarized in the remainder of this section.

The model uses the post-equalization SINR as a measure for the signal quality. For the spatial layer $n$, it is defined to be the quotient of the received signal power and the sum of interference and noise powers, i. e.

$$\tilde{\gamma}_n = \frac{|x_{b_u^\star,n}|^2}{\sum_{i\in\mathcal{I}_u}\sum_{m=1,\dots,N_i} c_{i,n,m}|x_{i,m}|^2 + \sum_{m=1,\dots,N_{\mathrm{Rx}}} g_{n,m}|n|^2}. \tag{5.7}$$

Here, $x_{b,n}$ is the $n$th element of the vector $\mathbf{x}_b$. $N_i$ is the number of spatial layers on which interferer $i$ transmits. The variable $c_{i,n,m}$ stands for the amplification of the interference caused by spatial layer $m$ of interferer $i$ and received on spatial layer $n$. It is the element $\mathbf{C}_i[n,m]$ of matrix

$\mathbf{C}_i = \mathbf{G}\mathbf{H}_{u,i}\mathbf{W}_i$. Similarly, $g_{n,m}$ represents the amplification of the noise received on antenna $m$ effecting layer $n$, and is equivalent to the element $\mathbf{G}[n, m]$.

To reduce computational effort during simulation, the implementation of this model is split into two components. In the *offline component* a trace of the small-scale fading channel model is calculated, processed further, and stored on disk. Here, locations of UEs, scheduling decisions, and interference are not yet known. The preprocessed data is read by the *online component*. Based on scheduling decisions and macro-scale attenuation effects known at simulation runtime, this calculates signal quality and derives capacity and decode probability. The model is designed such as most of the computational effort occurs in the offline component. This allows for efficient simulations, because a set of trace files generated by the offline component can be reused for many simulations.

The MIMO configuration selected to serve a UE is not known to the offline component. This configuration defines the number of virtual transmit antennas, the number of spatial streams, and the precoding matrix. One possible approach to solve this would be storing channel qualities for all possible combinations on disk. However, this is overly expensive. The model therefore assumes that the precoding matrix which delivers highest possible performance is independent of the signal attenuation.

The offline component considers each number of virtual transmit antennas and spatial streams. For each of these, it evaluates all entries from the respective precoding codebook. It selects that precoding matrix $\mathbf{W}_{b_u^\star}$ which delivers highest spectral efficiency for an expected SINR of 10 dB.[11] This results in different precoding matrices being used per PRB pair.

The model is further simplified here by assuming that interference does not depend on the MIMO configuration actually used by an interferer. Instead, interference is calculated using random MIMO configurations, i. e. random precoding matrices $\mathbf{W}_i$. Thus, interference in the simulation is only influenced by the transmit power used by the interferers. This is acceptable here, because we are not interested in joint optimization MIMO modes for interference reduction.

With these simplifications, for each number of virtual transmit antennas and spatial layers, the precoding matrices $\mathbf{W}_{b_u^\star}$ and $\mathbf{W}_i$ are known to the offline component. It can thus calculate the parameters $\tilde{c}_{i,n} = \sum_{m=1,\ldots,N_i} c_{i,n,m}$ and $\tilde{g}_n = \sum_{m=1,\ldots,N_{\text{Rx}}} g_{n,m}$ in equation (5.7). Note that these are calculated and stored for each number of virtual transmit antennas, each number of spatial layers, and each PRB pair.

When the online component is executed, the number of virtual transmit antennas and the number of spatial layers has been selected. The online component reads the respective values from the trace file and calculates $\tilde{\gamma}_n$ as defined in equation (5.7), considering the transmit powers of the serving and the interfering BSs. Separate values are calculated for each spatial layer and each PRB pair. Interference caused by CRSs from neighboring cells is not considered. Instead, a cell is assumed to cause interference only if data is transmitted on the respective PRBs pair.

Subsequently, values belonging to the same codeword are combined to an effective SINR. Thereto they are mapped to mutual information (known as mutual information effective SINR metric, MIESM), averaged, and mapped back [Bru+05]. As the mutual information depends on the

[11] For a discussion of the error introduced by this simplification refer to Colom Ikuno [Col13, appendix A].

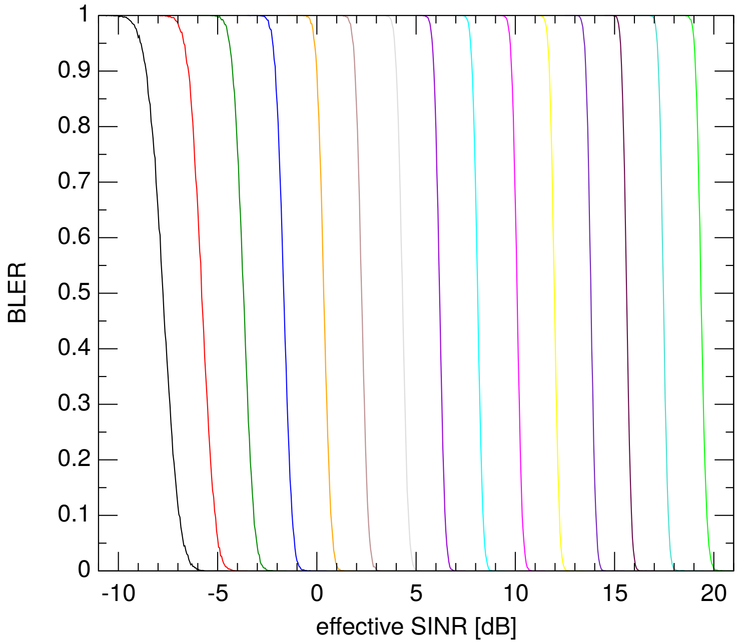


**Figure 5.3:** BLER curves used to model channel capacity (each curve corresponds to a combination of modulation and code rate as listed in table 2.5 on page 36; data from [RST16; Meh+11])

modulation scheme, three different effective SINR values are calculated for the three modulation schemes standardized in LTE.

The resulting effective SINR is taken as input for AWGN *block error rate* (BLER) curves. These are taken from the *Vienna LTE-A Downlink Link Level Simulator* [RST16; Meh+11]. For each modulation scheme and code rate in the CQI table (see table 2.5), a separate curve maps the effective SINR to a BLER. The BLER curves are plotted in figure 5.3. The optimal combination of modulation scheme and code rate can be selected by iterating over all combinations and selecting the one which maximizes the capacity. To avoid a high BLER, only those MCSs are considered which result in a decode probability above 80 %.

When calculating the capacity, overhead for CRSs transmitted on four antenna ports and for a control region with a fixed size of three OFDM symbols is taken into account. According to the description in section 2.4.2.1, there are also other sources for overhead, e. g. CSI-RSs, DM-RSs, and synchronization signals. These are not modeled here.

The decode probability of each codeword is taken from the same BLER tables. A Bernoulli experiment decides whether a codeword could be decoded by the receiver. A simplified model for ARQ and HARQ is applied, which just retransmits a non-decodable codeword after 8 ms. The number of retransmissions is not limited. Control channels and ARQ feedback are assumed to be ideal, i. e. there are no additional sources of error.

### 5.3.6 Data Traffic Models

Two different data traffic models are applied in this thesis. First, a full-buffer model is used for optimization and for comparative evaluation by simulation. Second, a dynamic traffic model is used for simulation only, to verify that the proposed system can cope with varying load.

**Table 5.2:** Parametrization of the traffic model [NGM08]

| Object type | $\sigma$ | $\mu$ | $x_s^{\text{min}}$/B | $x_s^{\text{max}}$/MiB |
|---|---|---|---|---|
| FTP | 0.35 | 14.45 | 1 | 5 |
| HTTP main | 1.37 | 8.37 | 100 | 2 |
| HTTP embedded | 2.36 | 6.17 | 50 | 2 |

In the full-buffer model, a fixed number of UEs is placed in the scenario. Each tries to download an infinite amount of data. The model resembles a high load situation, i. e. there are no resources in the system left free because they are not requested by any UE.[12] While it does not resemble the dynamic behavior of Internet traffic, it can be assumed valid when evaluating a snapshot in time. The full-buffer model is used for optimization studies, because it allows to formulate simpler problems.

Compared to a realistic traffic model, the combination of the uniform distribution of UEs and the full-buffer model leads to a distortion of the distribution of load over the serving area of a cell. This is caused by the BS serving the UEs with different rates. In a realistic model, sizes of data requests are limited. Therefore, a higher rate leads to an earlier satisfaction of a request. At the same time, UEs served with a lower rate stay longer in the system. This results in a non-uniform distribution of active UEs over the serving area, with a higher density in those regions where the average serving rate is lower. This effect is not captured by the full-buffer model.

The dynamic traffic model describes the arrival of objects on application layer at the BS for transmission in DL direction. It consists of an arrival process and an object size distribution. In addition, it defines the interaction of the traffic model with the UE placement.

The IAT of objects is modeled to be negative exponentially distributed. This resembles a large number of passive users which cause new requests randomly and independently of each other. The mean IAT is a parameter and is used to configure the average load of the network.

The sizes of the objects are modeled in accordance with the models for *hypertext transfer protocol* (HTTP) and *file transfer protocol* (FTP) traffic described by NGMN [NGM08]. The remaining application classes from that model are omitted, because it is assumed that HTTP and FTP introduce most of the dynamic, while other application classes like video and voice cause rather constant load. The ratio of HTTP and FTP traffic is the same as proposed by NGMN, i. e. ⅔ of the requests are HTTP requests and ⅓ are FTP requests.

While a FTP download consists of a single object, a HTTP transfer consists of a main website object and a number of embedded objects. These are here combined to a single object by summing up their sizes, so that modeling of the behavior of the web browsers of the UEs can be avoided. This resembles, e. g., the *Server Push* feature of HTTP/2.

[12] This assumes that there is no BS which serves no UE. Given that the total number of UEs in the system is high, such a situation occurs with very low probability. It was not observed during the studies for this thesis.

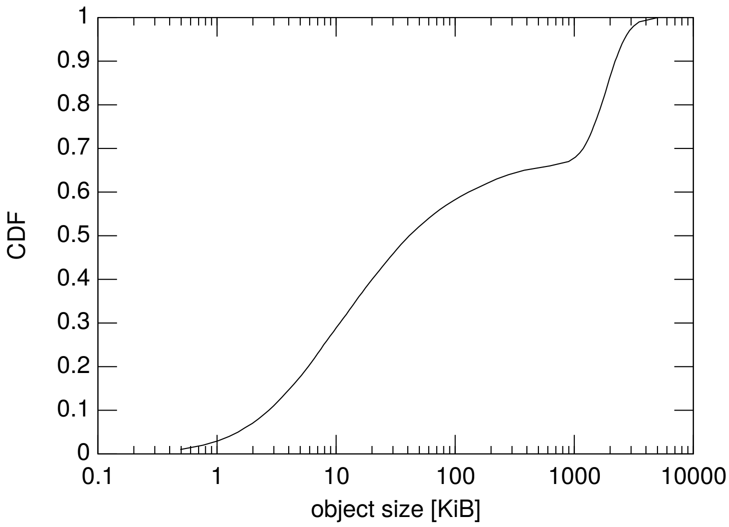


**Figure 5.4:** Object sizes of the traffic model

The object sizes are modeled as log-normally distributed, i. e. the *probability density function* (PDF) of the object size $x_s$ in bytes is

$$f(x_s) = \frac{1}{\sqrt{2\pi}\sigma x_s} e^{\frac{-(\ln x_s - \mu)^2}{2\sigma^2}} \quad \text{for } x_s > 0. \tag{5.8}$$

The parametrization is specified in table 5.2. The object size distributions are truncated, so that the resulting object sizes lie in the range of $x_s^{\min}$ to $x_s^{\max}$. The truncation is performed by re-sampling, i. e. when the size is outside the range, a new random number is drawn.

The number of embedded objects per web page, denoted as $x_n$, is determined by a modified Pareto distribution with the PDF

$$f(x_n) = \begin{cases} \frac{\alpha k^\alpha}{(x_n+k)^{\alpha+1}} & \text{for } 0 \le x_n < m - k \\ \left(\frac{k}{m}\right)^\alpha & \text{for } x_n = m - k \end{cases} \tag{5.9}$$

with the parametrization $\alpha = 1.1$, $k = 2$, and $m = 55$. Here, the second case (for $x_n = m - k$) corresponds to the total probability integrated over the cut off tail. The actual number of embedded objects is an integer value, therefore the resulting random number is rounded.[13]

The resulting HTTP pages have an average size of 56 KiB, which includes main and embedded objects. The average size of the FTP objects is 1.91 MiB. The overall average object size is 688 KiB. Figure 5.4 shows the CDF of the object sizes. The contributions of the HTTP and the FTP model are visible clearly.

Whenever the traffic model creates a new object, a new UE is placed at random coordinates as described in section 5.3.3. The serving BS is determined by evaluating the channel model. Then, a simple AC mechanism is executed:[14] Whenever the serving BS already has $N_{\mathrm{UE}}^{\max} = 100$

[13] Note that the specification of equation (5.9) in the original document [NGM08] is incomplete and partially broken. Other sources of the same or a closely related model ([Sri+08] and [3GPP2 09], respectively) have similar problems. The formula was here corrected to follow the PDF of a shifted Pareto distribution.

[14] See section 3.2.2 for a reasoning for AC.

UEs currently transmitting a data object, the newly arriving request is dropped. The remaining data objects are transmitted from the serving BS to the receiving UE. Thereby, it is assumed that arbitrary segmentation of the data objects is possible without any overhead. The object is successfully transmitted when all segments have arrived at the UE. Subsequently, the UE is removed from the simulation scenario.

### 5.3.7 Processing Effort Model

A model for the compute effort for signal processing is required, which captures the effect of RA and LA. We are here not interested in the raw number of calculations (e. g. number of *floating point operations*, FLOPs), but in the total time required by a C-RAN BBU to perform a certain calculation. This includes time for task switching, memory accesses, pipelining latencies, and other overhead. A bottom-up model, where the number of calculations for certain signal processing algorithms is counted, is therefore not adequate. Instead, a top-down model, which is based on measurements of a complete system, is used.

There are some publications where the computational effort of a software implementation of a BBU was measured [I+14a; Wei+12; Kai+12; Bha+12]. However, there is no information about how that effort scales with RA, modulation, coding, and MIMO configuration.[15] Therefore, in this thesis a model is used which is based on a publication by Desset et al. [Des+12]. Their model is originally intended for calculating the power consumption of a dedicated BBU, but provides numbers for computational effort as intermediate step.

The model calculates the effort in *giga operations per second* (GOPS). However, that does not encompass raw numerical calculations, only. Instead, the model states that hardware capable of performing operations with a certain rate (GOPS) is required for a task. It thus captures the overhead mentioned above.

Additional insight on this is provided by another publication of some of the same authors [DDL13]. That compares a bottom-up model for calculation of the FFT with the corresponding numbers from the top-down model. The results differ significantly, and the authors provide a list of sources for the differences. They conclude that the top-down model includes different sources of overhead. In this thesis, the focus is on the compute hardware required to operate a BBU. As that also has to cope with overhead, the top-down model is assumed to be suitable here.

The model defines the computational effort for the system components DPD, filtering, CPRI, OFDM, frequency-domain processing, and FEC. For each of the components, the effort is described as a function of the parameters bandwidth, number of antennas, modulation scheme, code rate, and time and frequency domain utilization of the system. This is realized by defining reference values for the parameters and scaling exponents.

Some of the components have a constant effort independent of RA or other dynamic decisions (see also section 5.1.2). These components are omitted from the derived model in this thesis,

[15] Bhaumik et al. [Bha+12] provide some insight on how the execution time of signal processing jobs scales with modulation, coding, and number of allocated resources. They state that load is a linear function of MCS and allocated PRBs, which is equivalent to the model used here. However, they did not investigate the influence of MIMO configuration.

leaving the components frequency-domain processing and FEC. Frequency-domain processing comprises modulation and MIMO operations. In the original model, it is split into a linear and a non-linear component. This results in the following formula for the processing effort caused by the whole system:

$$P_{\text{total}} = \sum_{c \in \mathcal{C}} P_{c,\text{ref}} \prod_{x \in \mathcal{X}} \left( \frac{x_{\text{act}}}{x_{\text{ref}}} \right)^{s_{c,x}}, \tag{5.10}$$

where $\mathcal{C}$ are the components and $\mathcal{X}$ the scaling parameters.

The reference effort $P_{c,\text{ref}}$ of component $c$, the reference values for the parameters $x_{\text{ref}}$, and the scaling exponents $s_{c,x}$ for all combinations of components $c$ and parameters $x$ are given in the publication. Inserting them into equation (5.10) results in

$$\begin{aligned} P_{\text{total}} &= 30 \left( \frac{b}{20\,\text{MHz}} \right) \left( \frac{a}{1} \right) \left( \frac{d_\text{t}}{1} \right) \left( \frac{d_\text{f}}{1} \right) \\ &+ 10 \left( \frac{b}{20\,\text{MHz}} \right) \left( \frac{a}{1} \right)^2 \left( \frac{d_\text{t}}{1} \right) \left( \frac{d_\text{f}}{1} \right) \\ &+ 20 \left( \frac{b}{20\,\text{MHz}} \right) \left( \frac{m}{6} \right) \left( \frac{c}{1} \right) \left( \frac{a}{1} \right) \left( \frac{d_\text{t}}{1} \right) \left( \frac{d_\text{f}}{1} \right) \qquad (5.11) \\ &= \frac{b}{20\,\text{MHz}} d_\text{t} d_\text{f} \left( 30a + 10a^2 + 20 \frac{m}{6} ca \right). \qquad (5.12) \end{aligned}$$

Here, $b$ stands for the system bandwidth, $a$ for the number of transmit antennas, $m$ for the modulation bits per symbol, $c$ for the code rate, and $d_\text{t}$ and $d_\text{f}$ for utilization in time and frequency domain, respectively. In equation (5.12), the three terms in the bracket represent the linear part of frequency domain processing, the non-linear part, and the FEC encoding, respectively.

Two modifications are applied to this model to make it better suitable for the evaluations performed here. First, the original model does not differentiate between the number of transmit antennas and the number of spatial streams. Instead, it is assumed that both are always equal. However, this is not the case in reality. Especially when the BSs are equipped with more antennas than the UEs, assuming a high number of spatial streams is not valid. Therefore, a separate parameter $l$ is introduced which describes the number of spatial streams.

That parameter is introduced into equation (5.12) as follows. FEC encoding depends on the actual amount of transmitted data. Consequently, the parameter $a$ is replaced by $l$ in the respective term. The non-linear frequency domain processing is assumed to comprise MIMO precoding operations. These consist mainly of matrix multiplications, which map the symbols from the spatial layers to the antennas. The complexity scales linearly with each dimension of the involved matrices. The same applies for the number of memory accesses and related overhead. Therefore, in the respective term of the model, $a^2$ is replaced with the product $al$. The described introduction of the additional parameter matches a later extension of the original model for large-scale antenna systems by the same authors [DDL14].

The second modification targets the granularity of the model. The original model is intended to describe the continuously generated effort of a system with a certain average load. However, we are here interested in the effort for each allocated resource. Consequently, the model is here scaled down to describe the effort that occurs when encoding a single PRB. Thereto, the parameters for system bandwidth and utilization are defined accordingly. The system bandwidth is here fixed to $b = 10\,\text{MHz}$. By configuring $d_\text{t} = \frac{1\,\text{ms}}{1\,\text{s}} = 0.001$, the utilization in time domain scales the resulting output from operations per second to operations per subframe. Similarly, $d_\text{f} = \frac{1}{50}$ accounts for the fact that a subframe here consists of $N_\text{PRB} = 50$ PRBs.

Applying both modifications to equation (5.12) results in the following model:

$$\begin{aligned} P_\text{PRB,giga} &= \frac{1}{2} 0.001 \frac{1}{50} \left(30a + 10al + 20\frac{m}{6}cl\right) \\ &= 10^{-4} \left(3a + al + \frac{1}{3}mcl\right) \end{aligned} \tag{5.13}$$

Here, $P_\text{PRB,giga}$ describes the compute effort to encode a single PRB in giga operations, i. e. $P_\text{PRB,giga} = 1$ corresponds to one billion operations. Scaling this to single operations results in

$$P_\text{PRB} = 10^5 \left(3a + al + \frac{1}{3}mcl\right) . \tag{5.14}$$

Based on equation (5.14), the total effort to transmit a subframe in a single cell is calculated by summing over the PRBs allocated to the associated UEs. Analogously, the effort for a set of pooled BBUs is derived by summing over all served cells.

The absolute numbers of operations are not relevant here, because only relative evaluations are performed in the following chapters. To avoid giving irrelevant information, all compute efforts are normalized. The normalization is performed such that 100 % correspond to the theoretical peak effort $P_n^\text{peak}$, where $n$ represents the total number of modeled PRBs, i. e. $n = |\mathcal{B}|N_\text{PRB}$. The value of $P_n^\text{peak}$ is calculated by evaluating equation (5.14) for the maximum values of each parameter. Precisely, when setting $a = 8$, $l = 4$, $m = 6$, and $c = 0.926$,[16] this results in $P_n^\text{peak} = n \cdot 6\,340\,800$.

The model used here is based on a single source in literature. For a rudimentary verification, appendix B provides a comparison of the model with measurement data from literature. Even if the numbers in the model are of questionable reliability, the nature of the modeled dependencies is reasonable. The absolute numbers resulting from this model are not of interest here. In case a more accurate model can be created which captures specific characteristics of a certain implementation, the methods applied in this thesis could be repeated for that scenario.

### 5.3.8 Resource Allocation Model

Section 2.4.3.1 described various aspects regarding the specification of allocated resources. Summarizing, in an LTE system it is not possible to allocate arbitrary PRBs. Instead, the eNodeB has to select one of multiple patterns to allocate VRBs, which are then mapped to PRBs by a predefined mapping function. For simplicity, this has not been modeled here. In

[16] Maximum code rate taken from table 2.5.

the model, localized mapping of VRBs to PRBs is used, so that a pair of VRBs is mapped to two consecutive PRBs on the same subcarriers. It is assumed that there are no restrictions regarding the combinations of allocated VRBs. As further simplification, in some configurations the optimization problem allows to allocate arbitrary fractions of VRBs to different UEs.

Following the introduction in section 2.4.3.2 and the discussion in section 2.4.4, the capacity of the control channels required to communicate RAs to UEs is limited. The actual signaling capacity in terms of number of DCIs depends on the patterns of allocated resources, the channel qualities, and the occurrence of overlapping search spaces. This capacity is shared between RAs for UL and DL transmissions and other control messages. The signaling restrictions are not modeled here. Instead, RA is not restricted by control channel capacity.

## 5.4 Optimization Problem

The optimization problem formulated in this section serves multiple purposes. First, it can be used to evaluate the maximum compute resource utilization in a high load scenario. Second, it serves to assess the potential to cope with limited compute resources without restricting to a certain implementation. It allows to determine the highest possible network performance under restricted compute capacity. Third, the problem can be used to identify the relevant variables which have to be adapted for high performance. Thereby, we can derive insights about efficient solution strategies for heuristics. Finally, a modified version of the optimization problem is used as a benchmark for the evaluation of the heuristic.

We are interested in the impact of limited compute capacity on network performance. According to the processing effort model introduced in section 5.3.7, the compute effort for the encoding of a transmission depends on the RA and on LA parameters. The relation of compute effort and network performance for LA parameters could be evaluated locally per cell. In contrast, a restriction of the RA, which results in time-frequency resources not being used for transmission, reduces interference in neighboring cells. This effect can only be studied in a multi-cell scenario.

The remainder of this section is structured as follows. First, section 5.4.1 classifies the approach chosen here w. r. t. the optimization problems introduced in sections 3.3 and 4.2.3. Section 5.4.2 describes how LA is modeled in the optimization problem and which simplifications have been applied. The basic problem is then specified in section 5.4.3. Subsequently, the problem is extended by different fairness formulations in section 5.4.4. Section 5.4.5 introduces variants of the basic problem which are used to analyze the influence of different adapted variables on the system efficiency. A summary is provided in section 5.4.6.

### 5.4.1 Approach and Classification

The optimization problem that is formulated and solved here represents a joint optimization of on-off power level IfCo, RA, and LA. The available compute capacity is formulated as an additional constraint.

The RA component of this problem resembles a rate adaptive RA problem as presented in section 3.3. Here, the power allocation is not adapted to the channel, but each BS transmits on each resource with either full power or not at all. Different objectives, corresponding to different fairness schemes from section 3.2.2, are evaluated.

The problem also performs IfCo, similar as those outlined in section 4.2.3. Here, we model a fixed number of power allocation resources with variable continuous size. Each of these resources has a different (fixed) combination of on-off power levels, so that all possible combinations are covered. The same channel realization is assumed for all resources. Continuously sized scheduling resources are allocated to UEs from arbitrary power allocation resources. To simplify the problem, we consider only the strongest interferers for interference calculation. This allows to precalculate the rates for all combinations of relevant interferers. Adaptation of MIMO configurations of neighboring BSs is not part of the problem. Therefore, for interference calculation, random MIMO modes are assumed to be used by the interferers, independent of the optimized LA parameters.

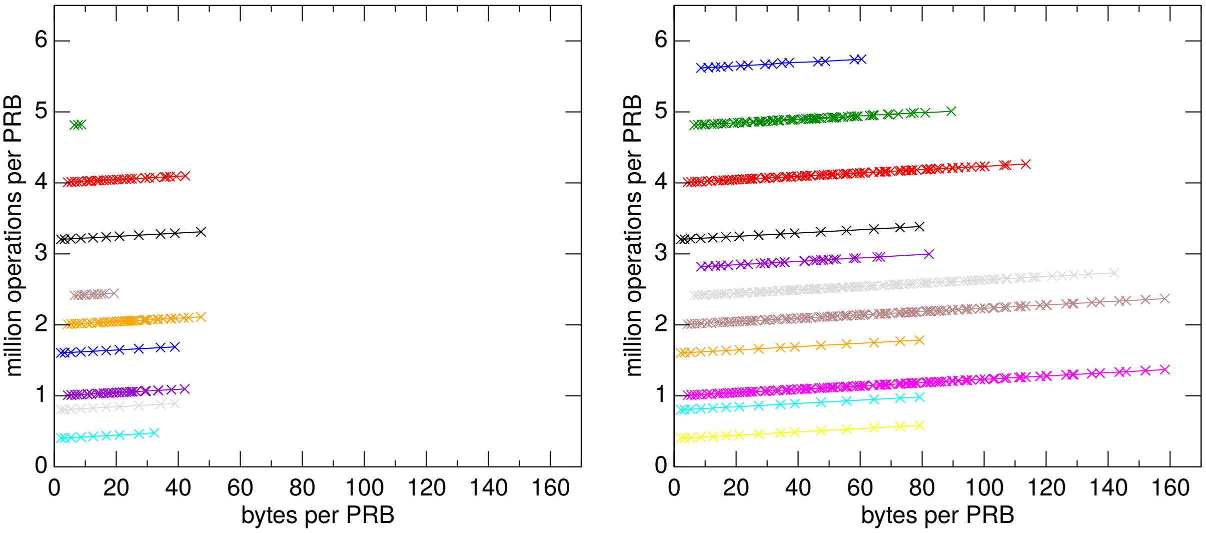
### 5.4.2 Modeling of Link Adaptation Parameters

The LA influences the parameters $a$, $l$, $m$, and $c$ of the processing effort model in equation (5.14). When LA is performed so that spectral efficiency is maximized, there is a jointly optimal configuration of these parameters. Configuring higher values for any of the parameters decreases the spectral efficiency and increases the compute effort, which is therefore not reasonable. In contrast, decreasing any of the values reduces both spectral efficiency and compute effort. It depends on the relation of both reductions whether it is efficient to perform such a configuration.

The parameters $a$ and $l$ resemble the MIMO mode of a transmission, and $m$ and $c$ the respective MCS. The MIMO mode parameters are contained in all terms of the processing effort equation (5.14), and appear even as a product on one of the terms. Therefore, reducing both results in a more-than-linear reduction of processing effort. At the same time, it is expected that the spectral efficiency does not scale linearly with the number of antennas and spatial streams, but is limited by the channel capacity. Therefore, adapting the MIMO mode is regarded a promising candidate for efficiently reducing processing effort. In contrast, the MCS parameters are contained only in one term of the processing effort equation (5.14). However, modulation and code rate have a linear influence on spectral efficiency. Therefore, the relation of reduction in spectral efficiency vs. reduction in compute effort is less favorable.

This is also illustrated by two examples in figure 5.5. The plots show two exemplary channel realizations, where the average wideband SINR is −0.1 dB and 17.4 dB, respectively. The spectral efficiency is plotted on the $x$-axis, the compute effort on the $y$-axis. Each point marked with × represents the combination of a MIMO mode and one or two MCSs.[17] For each given channel realization, this combination results in a certain spectral efficiency and compute effort, which determine the location of the respective point in the plot. Points standing for the same MIMO mode are drawn in the same color and are connected by a line. The actual MIMO modes, modulations, and code rates are of no relevance for the following discussion. Thus, they are not indicated. Points where the BLER for one of the codewords is below 80 % are not shown.

[17] For those MIMO modes which use two codewords, each point represents the combination of a MIMO mode and two MCSs.



**(a)** average wideband SINR −0.1 dB

**(b)** average wideband SINR 17.4 dB

**Figure 5.5:** Compute effort over spectral efficiency for two randomly drawn channel realizations

The slope of the curves is low. This means that, when changing modulation and coding, a small reduction of processing effort comes along with a large degradation of spectral efficiency. To evaluate the effect of changing MIMO modes, the rightmost points of the curves have to be compared. In contrast to changes in MCS, changes in MIMO mode result in stronger reductions of processing effort. At the same time, a high spectral efficiency can be maintained.

To simplify the optimization problem, parameters $m$ and $c$ are not modeled as variables of the optimization problem. Instead, only MIMO modes are optimized. For each MIMO mode, that MCS is applied which maximizes spectral efficiency. With that, spectral efficiency and compute effort can be derived. We allow simultaneous allocation of resources with different MIMO modes and therefore different MCSs, which is not possible with standardized LTE.

### 5.4.3 Mathematical Specification of the Basic Problem

The optimization problem encompasses the allocation of resources to UEs, its effect on the interference, and the selection of MIMO modes. The available compute resources, denoted as $p_{\text{max}}$, are a constraint for the optimization. Fairness and throughput are the base for the objective and, optionally, additional constraints. The problem shall be formulated by linear equations, so that it can be solved efficiently by standard solvers.[18]

#### *5.4.3.1 Resource Model and Basic Problem Formulation*

Each UE can be served with different MIMO modes. A MIMO mode is a feasible combination of a number of virtual transmit antennas $a$ and a number of spatial layers $l$. Virtual antennas are limited by the number of physical antennas which are installed per BS, i. e. $a \leq N_{\text{Tx}}$. The spatial layers are restricted by the number of receive antennas at the UE and the number of virtual

[18] Note that the later definitions also include a problem with binary variables, problem (5.41). However, that problem is simplified in other ways and can therefore also be solved efficiently.

transmit antennas, i. e. $l \le N_{\mathrm{Rx}}$ and $l \le a$. We assume the same number of transmit antennas $N_{\mathrm{Tx}}$ for all BSs and the same number of receive antennas $N_{\mathrm{Rx}}$ for all UEs. In principle, each UE can be served with any MIMO mode, although, depending on the channel, some modes will yield low throughput for some UEs. Therefore, there is a single set of MIMO modes usable by all nodes in the network, which is denoted as $\mathcal{M}$.

The resource model introduced in section 4.2.3 is extended by a third layer of resources. On all three layers of the resource model, the number of resources is fixed, and the sizes of the resources are variable.

As introduced before, we use on-off power levels for IfCo, only. To avoid unnecessary restriction of the usable reuse patterns, we allow all combinations of transmitting and non-transmitting BSs. Each combination is represented by a power allocation resource $n^{\mathrm{p}} \in \mathcal{R}^{\mathrm{p}}$. These resources are defined to be orthogonal, e. g. by separation in frequency domain. Assume that on a power allocation resource $n^{\mathrm{p}}$, $\mathcal{B}^{\mathrm{active}}_{n^{\mathrm{p}}} \subset \mathcal{B}$ are the transmitting BSs. Then, all possible sets of transmitting BSs, which are associated with the different power states, form the power set of $\mathcal{B}$, i. e. $\{\mathcal{B}^{\mathrm{active}}_{n^{\mathrm{p}}} : n^{\mathrm{p}} \in \mathcal{R}^{\mathrm{p}}\} = \mathcal{P}(\mathcal{B})$. This also determines the number of power allocation resources to be $|\mathcal{R}^{\mathrm{p}}| = |\mathcal{P}(\mathcal{B})| = 2^{|\mathcal{B}|}$.

A scheduling resource $n^{\mathrm{s}} \in \mathcal{R}^{\mathrm{s}}$ is defined for each combination of UE $u$ and power allocation resource $n^{\mathrm{p}}$, given that the UE's serving BS $b^{\star}_{u}$ is transmitting on the respective power allocation resource. This can be formulated as $\mathcal{R}^{\mathrm{s}} = \{n^{\mathrm{s}}_{u,n^{\mathrm{p}}} : u \in \mathcal{U}, n^{\mathrm{p}} \in \mathcal{R}^{\mathrm{p}}, b^{\star}_{u} \in \mathcal{B}^{\mathrm{active}}_{n^{\mathrm{p}}}\}$.

The assignment of UEs to BSs is modeled by grouping these scheduling resources per BS and power allocation resource:

$$\mathcal{R}^{\mathrm{s}}_{b,n^{\mathrm{p}}} = \left\{ n^{\mathrm{s}}_{u,\widetilde{n^{\mathrm{p}}}} \in \mathcal{R}^{\mathrm{s}} : b^{\star}_{u} = b, \widetilde{n^{\mathrm{p}}} = n^{\mathrm{p}} \right\} \quad \forall b \in \mathcal{B}, n^{\mathrm{p}} \in \mathcal{R}^{\mathrm{p}} \tag{5.15}$$

Thus, the set $\mathcal{R}^{\mathrm{s}}_{b,n^{\mathrm{p}}}$ contains the scheduling resources that are part of the same power allocation resource $n^{\mathrm{p}}$ and are dedicated to UEs which are served by BS $b$.

Scheduling resources inside such a group are assumed to be orthogonal. In addition, power allocation resources are orthogonal by definition. This ensures that in a single BS no resource is used twice. However, the same power allocation resource can be assigned to UEs served by a different BS. This other allocation can split the power allocation resource differently, i. e. use differently sized scheduling resources.

Each scheduling resource, which represents the allocation to one UE by its serving BS in one of the power allocation resources, is split into multiple *MIMO resources*. A MIMO resource $n^{\mathrm{m}} \in \mathcal{R}^{\mathrm{m}}$ is defined for each usable combination of power allocation resource, UE, and MIMO mode: $\mathcal{R}^{\mathrm{m}} = \{n^{\mathrm{m}}_{u,n^{\mathrm{p}},m} : u \in \mathcal{U}, n^{\mathrm{p}} \in \mathcal{R}^{\mathrm{p}}, b^{\star}_{u} \in \mathcal{B}^{\mathrm{active}}_{n^{\mathrm{p}}}, m \in \mathcal{M}\}$. MIMO resources are grouped by scheduling resource:

$$\mathcal{R}^{\mathrm{m}}_{n^{\mathrm{s}}_{u,n^{\mathrm{p}}}} = \left\{ n^{\mathrm{m}}_{\widetilde{u},\widetilde{n^{\mathrm{p}}},m} \in \mathcal{R}^{\mathrm{m}} : \widetilde{u} = u, \widetilde{n^{\mathrm{p}}} = n^{\mathrm{p}} \right\} \quad \forall n^{\mathrm{s}}_{u,n^{\mathrm{p}}} \in \mathcal{R}^{\mathrm{s}} \tag{5.16}$$

This means that the set $\mathcal{R}^{\mathrm{m}}_{n^{\mathrm{s}}_{u,n^{\mathrm{p}}}}$ contains the MIMO resources belonging to scheduling resource $n^{\mathrm{s}}_{u,n^{\mathrm{p}}}$, i. e., those which are dedicated to UE $u$ and part of the power allocation resource $n^{\mathrm{p}}$.

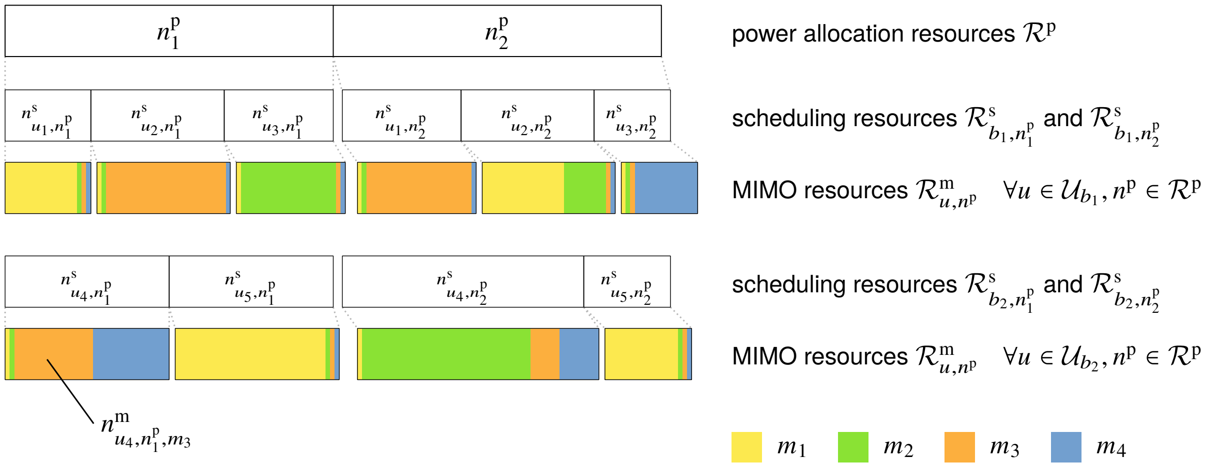

$n_1^\mathrm{p}$
$n_2^\mathrm{p}$
power allocation resources $\mathcal{R}^\mathrm{p}$
$n^\mathrm{s}_{u_1,n_1^\mathrm{p}}$
$n^\mathrm{s}_{u_2,n_1^\mathrm{p}}$
$n^\mathrm{s}_{u_3,n_1^\mathrm{p}}$
$n^\mathrm{s}_{u_1,n_2^\mathrm{p}}$
$n^\mathrm{s}_{u_2,n_2^\mathrm{p}}$
$n^\mathrm{s}_{u_3,n_2^\mathrm{p}}$
scheduling resources $\mathcal{R}^\mathrm{s}_{b_1,n_1^\mathrm{p}}$ and $\mathcal{R}^\mathrm{s}_{b_1,n_2^\mathrm{p}}$
MIMO resources $\mathcal{R}^\mathrm{m}_{u,n^\mathrm{p}}$ $\forall u \in \mathcal{U}_{b_1}, n^\mathrm{p} \in \mathcal{R}^\mathrm{p}$
$n^\mathrm{s}_{u_4,n_1^\mathrm{p}}$
$n^\mathrm{s}_{u_5,n_1^\mathrm{p}}$
$n^\mathrm{s}_{u_4,n_2^\mathrm{p}}$
$n^\mathrm{s}_{u_5,n_2^\mathrm{p}}$
scheduling resources $\mathcal{R}^\mathrm{s}_{b_2,n_1^\mathrm{p}}$ and $\mathcal{R}^\mathrm{s}_{b_2,n_2^\mathrm{p}}$
MIMO resources $\mathcal{R}^\mathrm{m}_{u,n^\mathrm{p}}$ $\forall u \in \mathcal{U}_{b_2}, n^\mathrm{p} \in \mathcal{R}^\mathrm{p}$
$n^\mathrm{m}_{u_4,n_1^\mathrm{p},m_3}$
$m_1$
$m_2$
$m_3$
$m_4$

**Figure 5.6:** Resource model used for the optimization problem

An exemplary split of the resources is shown in figure 5.6. The figure depicts two power allocation resources $n_1^\mathrm{p}$ and $n_2^\mathrm{p}$ with equal sizes. Two BSs $b_1$ and $b_2$ are active on both power allocation resources and serve three and two UEs, respectively.[19] The UEs get different amounts of resources allocated, which is shown by the different sizes of the scheduling resources.

For each UE, the assigned scheduling resource is split into four MIMO resources. Here, each color represents a different MIMO mode. The power allocation resources are orthogonal for the whole system. For each BS, the respective scheduling resources are orthogonal, as well as the MIMO resources. The orthogonality can be realized, e. g., by time or frequency multiplexing. Different BSs can transmit on the same resources, given that they are active on the respective power allocation resource.

For each MIMO resource $n^\mathrm{m}$, the channel characteristics of the respective UE are known. The received interference can be derived from the power allocation resource. Together with the MIMO mode, the MCS resulting in maximal spectral efficiency can be selected. This allows to calculate the spectral efficiency and compute effort.

We use $r_{n^\mathrm{m}}$ and $p_{n^\mathrm{m}}$ to denote the rate and effort per unit allocation, where a unit is defined to be a single PRB. Here, $r_{n^\mathrm{m}}$ is calculated according to the description in section 5.3.5. The value of $p_{n^\mathrm{m}}$ is derived from equation (5.14). The size of a MIMO resource $n^\mathrm{m}$ is denoted as $s_{n^\mathrm{m}}$. The rate $r_u$ of UE $u$ is then calculated as

$$r_u = \sum_{n^\mathrm{p} \in \mathcal{R}^\mathrm{p}} \sum_{n^\mathrm{m} \in \mathcal{R}^\mathrm{m}_{u,n^\mathrm{p}}} s_{n^\mathrm{m}} r_{n^\mathrm{m}}. \tag{5.17}$$

Based on these definitions, the following optimization problem can be formulated:

[19] Assume that there are other BSs and power allocation resources, so that on each power allocation resource a different subset of BSs is transmits.

$$
\begin{aligned}
\max_{\substack{s_{n^\mathrm{p}}, n^\mathrm{p} \in \mathcal{R}^\mathrm{p} \\ s_{n^\mathrm{s}}, n^\mathrm{s} \in \mathcal{R}^\mathrm{s} \\ s_{n^\mathrm{m}}, n^\mathrm{m} \in \mathcal{R}^\mathrm{m}}} \quad & \sum_{u \in \mathcal{U}} \mathrm{U}(r_u) && && (5.18\mathrm{a}) \\
\text{s. t.} \quad & \sum_{n^\mathrm{p} \in \mathcal{R}^\mathrm{p}} s_{n^\mathrm{p}} \le s_\mathrm{max} && && (5.18\mathrm{b}) \\
& \sum_{n^\mathrm{s} \in \mathcal{R}^\mathrm{s}_{b,n^\mathrm{p}}} s_{n^\mathrm{s}} \le s_{n^\mathrm{p}} && \forall b \in \mathcal{B}, n^\mathrm{p} \in \mathcal{R}^\mathrm{p} && (5.18\mathrm{c}) \\
& \sum_{n^\mathrm{m} \in \mathcal{R}^\mathrm{m}_{n^\mathrm{s}}} s_{n^\mathrm{m}} \le s_{n^\mathrm{s}} && \forall n^\mathrm{s} \in \mathcal{R}^\mathrm{s} && (5.18\mathrm{d}) \\
& \sum_{n^\mathrm{m} \in \mathcal{R}^\mathrm{m}} s_{n^\mathrm{m}} p_{n^\mathrm{m}} \le P^\mathrm{peak}_{|\mathcal{B}| s_\mathrm{max}} p_\mathrm{max} && && (5.18\mathrm{e})
\end{aligned}
$$

Here, the objective (5.18a) is to maximize the sum of the UEs’ utilities. The utility can in principle be different for each UE. However, this is not used here. Different utility functions will be defined in the following paragraphs to model different fairness schemes.

The first three constraints limit the sizes of the resources on the three levels of the resource model and thereby ensure orthogonality of resources inside each level. Equation (5.18b) ensures that not more than $s_\mathrm{max}$ resources are allocated. Equation (5.18c) models the fact that in each BS, each power allocation resource is split into non-overlapping scheduling resources. Equation (5.18d) formulates the same for the relation of scheduling and MIMO resources. Finally, the total compute effort is restricted by equation (5.18e). Here, $P^\mathrm{peak}_{|\mathcal{B}| s_\mathrm{max}}$ denotes the peak effort for this system as defined in section 5.3.7 and $p_\mathrm{max}$ is a parameter in the range 0 to 1.

#### *5.4.3.2 Simplification of Interference Relations*

The optimization problem in problem (5.18) is a linear problem with $N_\mathrm{var}$ continuous variables, where

$$
\begin{aligned}
N_\mathrm{var} &= |\mathcal{R}^\mathrm{p}| + |\mathcal{R}^\mathrm{s}| + |\mathcal{R}^\mathrm{m}| \\
&= |\mathcal{R}^\mathrm{p}| + \frac{1}{2}|\mathcal{R}^\mathrm{p}||\mathcal{U}| + \frac{1}{2}|\mathcal{R}^\mathrm{p}||\mathcal{U}||\mathcal{M}| \\
&= |\mathcal{R}^\mathrm{p}| \left(1 + \frac{1}{2}|\mathcal{U}| + \frac{1}{2}|\mathcal{U}||\mathcal{M}|\right) .
\end{aligned}
\tag{5.19}
$$

First, assume the following typical model configuration. The larger scenario introduced in section 5.3.2 has $|\mathcal{B}| = 57$ cells. When in average ten UEs are configured per cell, their total number is $|\mathcal{U}| = 570$. In addition, $|\mathcal{M}| = 11$ MIMO modes are assumed. These correspond to all possible numbers of spatial streams which can be transmitted with eight, four, two, and one virtual antennas. With these parameters, the problem has approximately $N_\mathrm{var} \approx 4.9 \cdot 10^{20}$ variables. This can be simplified by choosing a smaller scenario, e. g. with $|\mathcal{B}| = 21$ and

$|\mathcal{U}| = 210$. However, there the number of variables is $N_{\mathrm{var}} \approx 2.6 \cdot 10^9$, which is still difficult to solve.[20]

Therefore, the interference relations are simplified as follows. For each UE $u$, all BSs except the serving BS $b^\star_u$ can cause interference, i. e., they are the interferers $\mathcal{I}_u$ of $u$ with $\mathcal{I}_u = \mathcal{B} \setminus \{b^\star_u\}$. For the rate calculations of each UE, only the power states of the $N_{\mathrm{I}}$ strongest interferers are considered. These interferers are termed the *relevant interferers* $\mathcal{I}^{\mathrm{rel}}_u \subset \mathcal{I}_u$ of UE $u$. All other interferers are assumed to be transmitting on all power allocation resources. The relevant interferers of all UEs served by a BS $b$ are combined to $\mathcal{I}^{\mathrm{rel}}_b = \bigcup_{u \in \mathcal{U}_b} \mathcal{I}^{\mathrm{rel}}_u$, where $\mathcal{U}_b$ are the UEs served by BS $b$, i. e. $\mathcal{U}_b = \{u \in \mathcal{U} \colon b^\star_u = b\}$.

For all UEs of a BS, the power allocation to other BSs which are not in the set of relevant interferers does not need to be considered. Therefore, when considering the local resource allocation for this BS, the respective power allocation resources can be grouped, such that inside each group all power states of the relevant interferers are equal. The set of all groups of power allocation resources as seen by BS $b$ is here denoted as $\mathcal{R}^{\mathrm{g}}_b$. This is defined as

$$\mathcal{R}^{\mathrm{g}}_b = \left\{ n^{\mathrm{g}} \subset \mathcal{R}^{\mathrm{p}} \colon b \in \mathcal{B}^{\mathrm{active}}_{n^{\mathrm{p}}} \; \forall\, n^{\mathrm{p}} \in n^{\mathrm{g}}, \quad \mathcal{B}^{\mathrm{active}}_{n^{\mathrm{p}}_1} \cup \mathcal{I}^{\mathrm{rel}}_b = \mathcal{B}^{\mathrm{active}}_{n^{\mathrm{p}}_2} \cup \mathcal{I}^{\mathrm{rel}}_b \; \forall\, n^{\mathrm{p}}_1, n^{\mathrm{p}}_2 \in n^{\mathrm{g}} \right\} \qquad \forall b \in \mathcal{B} \tag{5.20}$$

From this, it follows that the union over all power allocation groups $\mathcal{R}^{\mathrm{g}}_b$ equals the set of those power allocation resources where the BS $b$ is transmitting, i. e. $\bigcup_{n^{\mathrm{g}} \in \mathcal{R}^{\mathrm{g}}_b} = \{n^{\mathrm{p}} \in \mathcal{R}^{\mathrm{p}} \colon b \in \mathcal{B}^{\mathrm{active}}_{n^{\mathrm{p}}}\}$. We define the size of a power allocation group $n^{\mathrm{g}}$ to be $s_{n^{\mathrm{g}}}$.

Scheduling resources can now be redefined such that there is one scheduling resource for each UE and power allocation group of the respective serving BS, i. e. $\mathcal{R}^{\mathrm{s}} = \{n^{\mathrm{s}}_{u,n^{\mathrm{g}}} \colon u \in \mathcal{U}, n^{\mathrm{g}} \in \mathcal{R}^{\mathrm{g}}_{b^\star_u}\}$. Scheduling resources are grouped by BS and power allocation group:

$$\mathcal{R}^{\mathrm{s}}_{b,n^{\mathrm{g}}} = \left\{ n^{\mathrm{s}}_{u,\widetilde{n^{\mathrm{g}}}} \in \mathcal{R}^{\mathrm{s}} \colon b^\star_u = b, \widetilde{n^{\mathrm{g}}} = n^{\mathrm{g}} \right\} \qquad \forall b \in \mathcal{B}, n^{\mathrm{g}} \in \mathcal{R}^{\mathrm{g}}_b \tag{5.21}$$

Similarly, MIMO resources are redefined such that there is one MIMO resource per power allocation group, UE, and MIMO mode: $\mathcal{R}^{\mathrm{m}} = \{n^{\mathrm{m}}_{u,n^{\mathrm{g}},m} \colon u \in \mathcal{U}, n^{\mathrm{g}} \in \mathcal{R}^{\mathrm{g}}_{b^\star_u}, m \in \mathcal{M}\}$. Again, MIMO resources are grouped by scheduling resource:

$$\mathcal{R}^{\mathrm{m}}_{n^{\mathrm{s}}_{u,n^{\mathrm{g}}}} = \left\{ n^{\mathrm{m}}_{\widetilde{u},\widetilde{n^{\mathrm{g}}},m} \colon \widetilde{u} = u, \widetilde{n^{\mathrm{g}}} = n^{\mathrm{g}} \right\} \qquad \forall n^{\mathrm{s}} \in \mathcal{R}^{\mathrm{s}} \tag{5.22}$$

With these definitions, the optimization problem (5.18) can be reformulated as

[20] Note that some of the variables in the model depend on each other. E. g., the size of the power allocation resources can be calculated by summing over the respective sets of scheduling resources. This allows for a limited simplification of the problem, however the complexity still scales with the product of $|\mathcal{R}^{\mathrm{p}}|$, $|\mathcal{U}|$, and $|\mathcal{M}|$. Such simplifications are not considered here.

$$\max_{\substack{s_{n^{\mathrm{p}}}, n^{\mathrm{p}} \in \mathcal{R}^{\mathrm{p}} \\ s_{n^{\mathrm{g}}}, n^{\mathrm{g}} \in \mathcal{R}^{\mathrm{g}}_b, b \in \mathcal{B} \\ s_{n^{\mathrm{s}}}, n^{\mathrm{s}} \in \mathcal{R}^{\mathrm{s}} \\ s_{n^{\mathrm{m}}}, n^{\mathrm{m}} \in \mathcal{R}^{\mathrm{m}}}} \quad \sum_{u \in \mathcal{U}} \mathrm{U}(r_u) \tag{5.23a}$$

$$\text{s. t.} \quad \sum_{n^{\mathrm{p}} \in \mathcal{R}^{\mathrm{p}}} s_{n^{\mathrm{p}}} \leq s_{\max} \tag{5.23b}$$

$$\sum_{n^{\mathrm{p}} \in n^{\mathrm{g}}} s_{n^{\mathrm{p}}} \geq s_{n^{\mathrm{g}}} \qquad \forall b \in \mathcal{B}, n^{\mathrm{g}} \in \mathcal{R}^{\mathrm{g}}_b \tag{5.23c}$$

$$\sum_{n^{\mathrm{s}} \in \mathcal{R}^{\mathrm{s}}_{b,n^{\mathrm{g}}}} s_{n^{\mathrm{s}}} \leq s_{n^{\mathrm{g}}} \qquad \forall b \in \mathcal{B}, n^{\mathrm{g}} \in \mathcal{R}^{\mathrm{g}}_b \tag{5.23d}$$

$$\sum_{n^{\mathrm{m}} \in \mathcal{R}^{\mathrm{m}}_{n^{\mathrm{s}}}} s_{n^{\mathrm{m}}} \leq s_{n^{\mathrm{s}}} \qquad \forall n^{\mathrm{s}} \in \mathcal{R}^{\mathrm{s}} \tag{5.23e}$$

$$\sum_{n^{\mathrm{m}} \in \mathcal{R}^{\mathrm{m}}} s_{n^{\mathrm{m}}} p_{n^{\mathrm{m}}} \leq P^{\mathrm{peak}}_{|\mathcal{B}| s_{\max}} p_{\max} \tag{5.23f}$$

Instead of equation (5.18c), the power allocation groups do here decouple the sizes of the scheduling resources from the sizes of the power allocation resources. Equation (5.23c) limits the size of each power allocation group to the sum of the sizes of the contained power allocation resources. Then, equation (5.23d) aligns the scheduling resources in each BS to the power allocation groups of that BS.

The number of power allocation groups is not a parameter but depends on the interference relations between BSs and UEs. The UEs associated to a BS see similar interferers. With increasing number of UEs, the relevant interferers overlap, and $\mathcal{I}^{\mathrm{rel}}_b$ contains the neighbors of BS $b$. The number of power allocation groups per BS in general depends on the geographical dependencies of the BSs, but does not scale with the number of UEs. In evaluations with $N_{\mathrm{I}} = 3$, the average number of relevant interferers per BS was $\mathbb{E}_{\mathrm{b}}\left[\left|\mathcal{I}^{\mathrm{rel}}_b\right|\right] \approx 6.9$. Here, $\mathbb{E}_{\mathrm{b}}[\bullet]$ denotes the arithmetic mean evaluated over all $b \in \mathcal{B}$. In the same evaluation, the resulting average number of power allocation groups per BS was $\mathbb{E}_{\mathrm{b}}\left[\left|\mathcal{R}^{\mathrm{g}}_b\right|\right] \approx 189$.

Introducing power allocation groups reduces the numbers of scheduling and MIMO resources. The number of variables of the simplified optimization problem (5.23) can be estimated as

$$\begin{aligned} N^{\mathrm{simp}}_{\mathrm{var}} &= |\mathcal{R}^{\mathrm{p}}| + \sum_{b \in \mathcal{B}} \left|\mathcal{R}^{\mathrm{g}}_b\right| + |\mathcal{R}^{\mathrm{s}}| + |\mathcal{R}^{\mathrm{m}}| \\ &\approx |\mathcal{R}^{\mathrm{p}}| + \mathbb{E}_{\mathrm{b}}\left[\left|\mathcal{R}^{\mathrm{g}}_b\right|\right] |\mathcal{B}| + \mathbb{E}_{\mathrm{b}}\left[\left|\mathcal{R}^{\mathrm{g}}_b\right|\right] |\mathcal{U}| + \mathbb{E}_{\mathrm{b}}\left[\left|\mathcal{R}^{\mathrm{g}}_b\right|\right] |\mathcal{U}||\mathcal{M}| \\ &= |\mathcal{R}^{\mathrm{p}}| + \mathbb{E}_{\mathrm{b}}\left[\left|\mathcal{R}^{\mathrm{g}}_b\right|\right] \left(|\mathcal{B}| + |\mathcal{U}| + |\mathcal{U}||\mathcal{M}|\right) \end{aligned} \tag{5.24}$$

With the same small scenario parameters as above ($|\mathcal{B}| = 21$, $|\mathcal{U}| = 210$, and $|\mathcal{M}| = 11$), this results in $N^{\mathrm{simp}}_{\mathrm{var}} \approx 2.6 \cdot 10^6$. Compared to the non-simplified model, the simplification has reduced the number of variables by factor 1000. This simplification showed to be sufficient to solve the problem using standard software and with reasonable memory and compute capabilities.

### 5.4.4 Formulation of Fairness Requirements

Different fairness schemes were introduced in section 3.2.2. The desired fairness has a strong influence on IfCo and RA, which both affect the compute effort. Therefore, in the following sections multiple of those fairness schemes are incorporated into the optimization problem (5.23). Most can be realized by defining the utility function $\mathrm{U}(\bullet)$ and / or by formulating additional constraints. The fairness criterion based on normalized throughput and the CDF of the users' rates is omitted here, because no efficient formulation for the optimization problem was found.

#### *5.4.4.1 No Fairness*

The simplest version of the optimization problem is that which does not consider fairness. Instead, it maximizes the total system throughput. This can be realized by defining the utility function to equal the rate itself:

$$\mathrm{U}(r) := r \tag{5.25}$$

The system configuration without fairness is here denoted as *None*.

#### *5.4.4.2 Proportional Fairness*

The *PF* scheme can be realized by defining the utility function to be the logarithm of the rate. Without losing generality, we do here normalize the rate, and thus define the utility as

$$\mathrm{U}(r) := \log\left(\frac{r}{r^{\max}}\right), \tag{5.26}$$

where $r^{\max}$ is the maximum rate that can be realized by a single UE. Note that the normalization causes the resulting utility values to be negative. Similar to the normalization, the base of the logarithm is of no relevance for the outcome of the optimization, but only for the absolute values of the resulting utility. The natural logarithm is used in the implementation.

As stated previously, it is here desired to formulate a linear optimization problem. Therefore, the logarithm is approximated by a piecewise linear function. The normalization in equation (5.26) limits the input values to the interval $[0, 1]$. Low values are not relevant for optimization, because $\lim_{r\to 0} \mathrm{U}(r) = -\infty$. Therefore, such values do not need to be represented accurately.

Assume the relevant interval of input values is given by $[e^{u_{\min}}, 1]$, which results in outputs from the interval $[u_{\min}, 0]$. We want an equally accurate approximation over this interval. Therefore, the following approximation is based on equidistant steps on the utility axis. The interval $[u_{\min}, 0]$ is divided into $N_{\text{pieces}}$ subintervals, each covering $[u_{\min} + iu_{\text{piece}}, u_{\min} + (i + 1)u_{\text{piece}})$

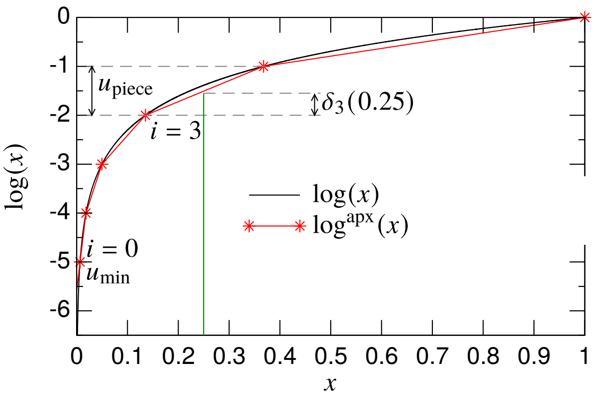


**Figure 5.7:** Example for the piecewise linear approximation with $N_{\text{pieces}} = 5$ and $u_{\min} = -5$

on the utility axis. Here, index $i$, with $i \in \{0, \ldots, N_{\text{pieces}} - 1\}$, identifies the subinterval, and $u_{\text{piece}}$, with $u_{\text{piece}} = \frac{-u_{\min}}{N_{\text{pieces}}}$, the size of each subinterval. Based on these definitions, the piecewise linear function is defined as

$$\log^{\text{apx}}(x) := \begin{cases} u_{\min} + \delta_0(x) & \text{for } \log(x) < u_{\min} \\ u_{\min} + i u_{\text{piece}} + \delta_i(x) & \text{for } u_{\min} + i u_{\text{piece}} \leq \log(x) < u_{\min} + (i+1) u_{\text{piece}} \\ & \quad \text{and } i \in \{0, \ldots, N_{\text{pieces}} - 1\} \\ \text{undefined} & \text{for } \log(x) \geq 0. \end{cases} \tag{5.27}$$

The function $\delta_i(x)$ represents the offset inside piece $i$ and is defined as

$$\begin{aligned} \delta_i(x) &:= u_{\text{piece}} \frac{x - e^{u_{\min} + i u_{\text{piece}}}}{e^{u_{\min} + (i+1) u_{\text{piece}}} - e^{u_{\min} + i u_{\text{piece}}}} \\ &= x \frac{u_{\text{piece}}}{e^{u_{\min} + (i+1) u_{\text{piece}}} - e^{u_{\min} + i u_{\text{piece}}}} - \frac{u_{\text{piece}} e^{u_{\min} + i u_{\text{piece}}}}{e^{u_{\min} + (i+1) u_{\text{piece}}} - e^{u_{\min} + i u_{\text{piece}}}}. \end{aligned} \tag{5.28}$$

Note that, for values below $u_{\min}$, the slope of the lowest piece is continued. This overestimates the utility for $r < r^{\max} e^{u_{\min}}$. Compared to the exact utility function, optimization with the approximated function could tend to favor this range. To avoid inaccuracies caused by a misconfiguration of $u_{\min}$, the optimal rates are checked after each optimization run. Results are not used in case the less accurate range of the utility function is used.

Figure 5.7 depicts an example for the approximation with $\log^{\text{apx}}(x)$. In this example, the linear pieces are defined by the two parameters $N_{\text{pieces}} = 5$ and $u_{\min} = -5$. Here, the value $x = 0.25$ lies in the range between $i = 3$ and $i = 4$. It is therefore approximated as $\log^{\text{apx}}(0.25) = u_{\min} + 3u_{\text{piece}} + \delta_3(0.25) = -5 + 3 + 0.49 = -1.51$.

The function $\log^{\text{apx}}(x)$ is concave. Therefore, in the maximization in problem (5.23) it can be represented as a set of inequalities. Assume that a set of additional variables $r_u^{\log}$ with $u \in \mathcal{U}$ is defined. Then, the optimization problem is

$$\max_{\substack{s_{n^\mathrm{p}}, n^\mathrm{p} \in \mathcal{R}^\mathrm{p} \\ s_{n^\mathrm{g}}, n^\mathrm{g} \in \mathcal{R}^\mathrm{g}_b, b \in \mathcal{B} \\ s_{n^\mathrm{s}}, n^\mathrm{s} \in \mathcal{R}^\mathrm{s} \\ s_{n^\mathrm{m}}, n^\mathrm{m} \in \mathcal{R}^\mathrm{m} \\ r_u^\mathrm{log}, u \in \mathcal{U}}} \quad \sum_{u \in \mathcal{U}} r_u^\mathrm{log} \tag{5.29a}$$

$$\text{s. t.} \quad r_u^\mathrm{log} \leq u_\mathrm{min} + i u_\mathrm{piece} + \delta_i \left( \frac{r_u}{r^\mathrm{max}} \right) \quad \forall u \in \mathcal{U}, i \in \{0, \ldots, N_\mathrm{pieces} - 1\} \tag{5.29b}$$

and equations (5.23b), (5.23c), (5.23d), (5.23e), and (5.23f).

#### 5.4.4.3 Guaranteed Minimum Rates

A simpler formulation of fairness can be achieved when a minimum rate is guaranteed for each UE. This scheme is here denoted as *MinR*. Assume that the minimum rate is represented by the parameter $r^\mathrm{min}$. The utility is defined to maximize the sum throughput as in equation (5.25). The optimization problem is than that defined in problem (5.23), with the additional constraint

$$r_u \geq r^\mathrm{min} \quad \forall u \in \mathcal{U}. \tag{5.30}$$

#### 5.4.4.4 Max-Min Fairness

The objective of the *MaxMin* fairness scheme is to equally maximize the rate of the UE with the lowest rate. This can be formulated by defining a single variable $r^\mathrm{min}$, which represents the rate of the UE with the lowest rate. This variable also resembles the objective function. The resulting problem is

$$\max_{\substack{s_{n^\mathrm{p}}, n^\mathrm{p} \in \mathcal{R}^\mathrm{p} \\ s_{n^\mathrm{g}}, n^\mathrm{g} \in \mathcal{R}^\mathrm{g}_b, b \in \mathcal{B} \\ s_{n^\mathrm{s}}, n^\mathrm{s} \in \mathcal{R}^\mathrm{s} \\ s_{n^\mathrm{m}}, n^\mathrm{m} \in \mathcal{R}^\mathrm{m} \\ r^\mathrm{min}}} \quad r^\mathrm{min} \tag{5.31a}$$

$$\text{s. t.} \quad r_u \geq r^\mathrm{min} \quad \forall u \in \mathcal{U} \tag{5.31b}$$

and equations (5.23b), (5.23c),

(5.23d), (5.23e), and (5.23f).

As discussed in section 3.2.2, there can be UEs that can receive a higher rate than the optimal $r^\mathrm{min}$. The formulation in problem (5.31) does not specify how to treat these UEs. Consequently, the resulting allocations could depend on internals of the optimizer. The formulation is changed, so that this problem is avoided.

In section 3.2.2, an extended definition was presented. This could be applied recursively by fixing the rates which cannot be increased further and then re-solving the problem for the remaining UEs. However, this cannot be formulated in a single linear program, and recursively solving the problem would result in high computational effort.

One viable approach is to enforce equality of all rates. However, that would result in low resource usage, because a single UE with unfavorable channel conditions prohibits allocation of free resources to other UEs. To avoid that, the sum throughput of the remaining UEs is maximized here. This is added to the utility function with a low scaling factor $\epsilon \ll 1$, so that the problem can be solved in one step:

$$\begin{aligned} \max_{\substack{s_{n^\mathrm{p}}, n^\mathrm{p} \in \mathcal{R}^\mathrm{p} \\ s_{n^\mathrm{g}}, n^\mathrm{g} \in \mathcal{R}^\mathrm{g}_b, b \in \mathcal{B} \\ s_{n^\mathrm{s}}, n^\mathrm{s} \in \mathcal{R}^\mathrm{s} \\ s_{n^\mathrm{m}}, n^\mathrm{m} \in \mathcal{R}^\mathrm{m} \\ r^\mathrm{min}}} \quad & r^\mathrm{min} + \epsilon \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} r_u \\ \text{s. t.} \quad & \text{equations (5.31b), (5.23b), (5.23c),} \\ & \text{(5.23d), (5.23e), and (5.23f)} \end{aligned} \tag{5.32}$$

Note that this results in an inaccuracy in the maximized minimum rate, as lower $r^\mathrm{min}$ can be traded for an increase in sum throughput. This effect has to be limited by choosing a sufficiently small value for the parameter $\epsilon$.

### 5.4.5 Restriction of Variables Adapted to Available Compute Resources

The optimization problem presented in section 5.4.3 jointly adapts on-off power level IfCo, RA, and LA to the available compute capacity. While this all-encompassing approach serves well for a first estimation of the possible efficiency of a compute effort reduction, it is difficult to implement in real systems. There, IfCo, RA, and LA are separate modules. Each of these modules already has a certain complexity. Therefore, we are interested in an approach that preferably consists of a simple modification of only one of these modules.

In this section, the optimization problem specified in section 5.4.3 is modified. Each modification resembles the adaptation of one of the above mentioned modules to the available compute capacity, while the remaining modules are ignorant of compute resources. The basis for the modifications is the assumption that IfCo, RA, and LA are performed in this order, and each of the steps bases its decisions on the outcome of the previous steps. It is further assumed that the final encoding of the data, which itself raises the compute effort, is executed as fourth step.

When a step is adapted to reduce compute effort, the decisions of the subsequent steps, which also have influence on the compute effort, have to be predicted. While this is an additional challenge for the implementation, the following modified optimization problems assume perfect prediction. This serves as an optimistic estimation for the performance of the adaptations. For the first three steps, sophisticated algorithms can be designed which compensate reduced compute capacity. In contrast, the last step can only downsize existing sets of allocated resources or skip their encoding.

#### *5.4.5.1 Adapting Interference Coordination*

When an IfCo algorithm is adapted to take compute capacity into account, it has to predict the compute effort related decisions of RA and LA. This is difficult to formulate as optimization problem, because the optimal resource allocation can be seen as an embedded problem with different constraints. This embedded problem would have to be re-solved for each possible configuration of the IfCo. Therefore, the performance of adapting IfCo is here approximated optimistically by assuming that IfCo as well as RA take compute capacity into account. Thus, only the LA does not consider compute capacity. Under these conditions, LA is a local problem, where it is always optimal to maximize spectral efficiency.

This results in only a single modification to the optimization problem (5.23) and the derived fairness-specific problems. There, MIMO resources model the freedom to choose different MIMO modes for the same link characteristics. This is here restricted, such that only the MIMO resource which maximizes spectral efficiency can be used.

Let $m^{\star}_{u,n^{\mathrm{g}}}$ denote the MIMO mode which results in the highest spectral efficiency for UE $u$ on a resource in power allocation group $n^{\mathrm{g}}$:

$$m^{\star}_{u,n^{\mathrm{g}}} = \operatorname*{arg\,max}_{m \in \mathcal{M}} r_{n^{\mathrm{m}}_{u,n^{\mathrm{g}},m}} \qquad \forall u \in \mathcal{U}, n^{\mathrm{g}} \in \mathcal{R}^{\mathrm{g}}_{b^{\star}_{u}} \tag{5.35}$$

Then, constraint (5.23e) in the original problem can be replaced by a fixed assignment, resulting in the following problem:

$$\max_{\substack{s_{n^{\mathrm{p}}}, n^{\mathrm{p}} \in \mathcal{R}^{\mathrm{p}} \\ s_{n^{\mathrm{g}}}, n^{\mathrm{g}} \in \mathcal{R}^{\mathrm{g}}_{b}, b \in \mathcal{B} \\ s_{n^{\mathrm{s}}}, n^{\mathrm{s}} \in \mathcal{R}^{\mathrm{s}}}} \quad \sum_{u \in \mathcal{U}} \mathrm{U}(r_u) \tag{5.36a}$$

$$\text{s. t.} \quad s_{n^{\mathrm{m}}_{u,n^{\mathrm{g}},m}} = \begin{cases} s_{n^{\mathrm{s}}_{u,n^{\mathrm{g}}}} & \text{for } m = m^{\star}_{u,n^{\mathrm{g}}} \\ 0 & \text{otherwise} \end{cases} \qquad \forall u \in \mathcal{U}, n^{\mathrm{g}} \in \mathcal{R}^{\mathrm{g}}_{b^{\star}_{u}}, m \in \mathcal{M} \tag{5.36b}$$

and equations (5.23b), (5.23c), (5.23d), and (5.23f)

Note that this allows further simplification of the problem, as it is not necessary to differentiate between scheduling resources and MIMO resources any more. This simplification is not further discussed here. The same modification can be applied to all derived fairness-specific problems.

#### *5.4.5.2 Adapting Resource Allocation*

To model the adaptation of RA alone, IfCo and LA are assumed to be optimized without considering the available compute capacity. For LA, this results in local maximization of spectral efficiency, as discussed before. However, for IfCo a separate optimization has to be performed. This IfCo problem is, in contrast to RA discussed in section 5.4.5.1, not an embedded problem. Instead, it can be solved beforehand, because it is not influenced by the RA.

Assume that the problem (5.23) or the derived fairness-specific problem has been solved for $p_{\max} = \infty$, or, equivalently, without constraint (5.23f). Let $\widehat{s_{n^{\mathrm{p}}}}$ denote the size of power allocation resource $n^{\mathrm{p}}$ in the solution of that problem. Define $m^{\star}_{u,n^{\mathrm{g}}}$ as in equation (5.35). Then, fix the sizes of the power allocation resources to those values. Note that this also allows to fix the sizes of the power allocation groups. In addition, constrain LA to always chose the mode for maximum spectral efficiency. The resulting problem is:

$$\max_{\substack{s_{n^\mathrm{s}}, n^\mathrm{s} \in \mathcal{R}^\mathrm{s} \\ s_{n^\mathrm{m}}, n^\mathrm{m} \in \mathcal{R}^\mathrm{m}}} \quad \sum_{u \in \mathcal{U}} \mathrm{U}(r_u) \tag{5.37a}$$

$$\text{s. t.} \quad s_{n^\mathrm{p}} = \widehat{s_{n^\mathrm{p}}} \qquad \forall n^\mathrm{p} \in \mathcal{R}^\mathrm{p} \tag{5.37b}$$

$$\text{and equations (5.36b), (5.23c), (5.23d), and (5.23f)}$$

Here, constraint (5.23d) allows to allocate less than the available resource size to each UE. However, as the sizes of the power allocation resources are fixed, this does not improve the interference conditions seen by the other UEs in the network. This models the fact that free resources, when not being coordinated with neighboring cells, cannot be used efficiently. The effect of statistically reduced interference is not considered here. The same modification can be applied to all derived fairness-specific problems.

#### *5.4.5.3 Adapting Link Adaptation*

Adapting only LA parameters (i. e., the MIMO mode) to the available compute capacity can be performed similarly. The optimal fairness-specific IfCo and RA have to be derived by solving a separate problem beforehand. These are then fixed, so that the sizes of the MIMO resources are the only variables left for the optimization.

There are two different approaches to adapt LA. First, a problem is formulated which comprises individual adaptation of parameters for each UE. Assume that the problem (5.23) (or the derived fairness-specific problem) has been solved for $p_\mathrm{max} = \infty$. Let $\widehat{s_{n^\mathrm{s}}}$ denote the size of scheduling resource $n^\mathrm{s}$ in the solution of that problem. Then, fix the sizes of the scheduling resources to those values. Note that this also renders the sizes of power allocation resources and power allocation groups irrelevant. The resulting problem is:

$$\max_{s_{n^\mathrm{m}}, n^\mathrm{m} \in \mathcal{R}^\mathrm{m}} \quad \sum_{u \in \mathcal{U}} \mathrm{U}(r_u) \tag{5.38a}$$

$$\text{s. t.} \quad s_{n^\mathrm{s}} = \widehat{s_{n^\mathrm{s}}} \qquad \forall n^\mathrm{s} \in \mathcal{R}^\mathrm{s} \tag{5.38b}$$

$$\text{and equations (5.23e) and (5.23f)}$$

Again, the same modification can be applied to all derived fairness-specific problems.

The second approach is to only restrict the number of virtual transmit antennas used by each cell, while the actual MIMO modes used are selected without considering the available compute resources. This simplifies the problem by defining only a single decision variable per cell.

Let $\mathcal{A}$ denote the set of supported configurations of transmit antennas. For each configuration $a \in \mathcal{A}$, $\mathcal{M}_a \subset \mathcal{M}$ represents the associated subset of allowed MIMO modes.[21] For each UE $u$,

[21] Here, a configuration of transmit antennas corresponds to a number of virtual antennas, e. g. $\mathcal{A} = \{1, 2, 4, 8\}$. The set of supported modes for a number of transmit antennas comprises those modes which do not rely on a larger number of transmit antennas. Thus, deliberately reducing the number of transmit antennas is allowed. This configuration is used such that the resulting model yields results comparable to the other problems, which also allow a deliberate reduction of transmit antennas.

power allocation resource $n^{\mathrm{g}}$, and configuration $a$, let $m^{\star}_{u,n^{\mathrm{g}},a}$ denote that MIMO mode which delivers the highest throughput, i. e.

$$m^{\star}_{u,n^{\mathrm{g}},a} = \underset{m \in \mathcal{M}_a}{\arg\max}\, r_{n^{\mathrm{m}}_{u,n^{\mathrm{g}},m}} \qquad \forall u \in \mathcal{U}, n^{\mathrm{g}} \in \mathcal{R}^{\mathrm{g}}_{b^{\star}_u}, a \in \mathcal{A}. \tag{5.39}$$

Based on that, introduce a flag $y_{m,u,n^{\mathrm{g}},a}$ which is one only if the mode $m$ delivers the highest throughput for the combination of the remaining indices, i. e.

$$y_{m,u,n^{\mathrm{g}},a} = \begin{cases} 1 & \text{for } m = m^{\star}_{u,n^{\mathrm{g}},a} \\ 0 & \text{otherwise} \end{cases} \qquad \forall u \in \mathcal{U}, n^{\mathrm{g}} \in \mathcal{R}^{\mathrm{g}}_{b^{\star}_u}, m \in \mathcal{M}, a \in \mathcal{A}. \tag{5.40}$$

Further, define a set of binary variables $x_{b,a}$ with $b \in \mathcal{B}$ and $a \in \mathcal{A}$. Each variable $x_{b,a}$ is equal to one if BS $b$ uses configuration $a$ and zero otherwise. As before, assume that the problem (5.23) has been solved for $p_{\max} = \infty$, and let $\widehat{s_{n^{\mathrm{s}}}}$ denote the size of scheduling resource $n^{\mathrm{s}}$ in the solution of that problem. With these definitions, the problem to cope with limited compute resources by selecting the number of transmit antennas per cell is:

$$\max_{x_{b,a}, b \in \mathcal{B}, a \in \mathcal{A}} \quad \sum_{u \in \mathcal{U}} \mathrm{U}(r_u) \tag{5.41a}$$

$$\text{s. t.} \quad s_{n^{\mathrm{m}}_{u,n^{\mathrm{g}},m}} = \widehat{s_{n^{\mathrm{s}}_{u,n^{\mathrm{g}}}}} \sum_{a \in \mathcal{A}} x_{a,b^{\star}_u}\, y_{m,u,n^{\mathrm{g}},a} \qquad \forall u \in \mathcal{U}, n^{\mathrm{g}} \in \mathcal{R}^{\mathrm{g}}_{b^{\star}_u}, m \in \mathcal{M} \tag{5.41b}$$

$$\sum_{a \in \mathcal{A}} x_{b,a} = 1 \qquad \forall b \in \mathcal{B} \tag{5.41c}$$

$$x_{b,a} \in \{0, 1\} \qquad \forall b \in \mathcal{B}, a \in \mathcal{A} \tag{5.41d}$$

and equation (5.23f)

Here, equation (5.41b) enforces the size of a MIMO resource to be equal to the size of the respective scheduling resource if and only if the associated MIMO mode is the one delivering maximal throughput for the currently active number of virtual transmit antennas. The two remaining constraints ensure that each BS uses only a single configuration for its transmit antennas. As before, the same modification can be applied to all derived fairness-specific problems.

#### *5.4.5.4 Downsizing Existing Sets of Allocated Resources*

Downsizing of existing RA, without modifying any other parameters, can be seen as last resort to cope with limited compute capacity. As before, optimal IfCo, RA, and LA parameters are derived from a separate optimization problem solved for unrestricted compute capacity.[22] These do then constraint the main problem, so allocated resources can only be decreased.

Assume that the problem (5.23) (or the derived fairness-specific problem) has been solved for $p_{\max} = \infty$, and let $\widehat{s_{n^{\mathrm{m}}}}$ denote the size of MIMO resource $n^{\mathrm{m}}$ in the solution of that problem.

[22] Note that this results in LA to be configured to maximize spectral efficiency. The same could have been achieved by applying constraint (5.36b).

Then, limit the sizes of the MIMO resources to not exceed those values. The resulting problem is:

$$\max_{s_{n^{\mathrm{m}}}, n^{\mathrm{m}} \in \mathcal{R}^{\mathrm{m}}} \quad \sum_{u \in \mathcal{U}} \mathrm{U}(r_u) \tag{5.42a}$$

$$\text{s. t.} \quad s_{n^{\mathrm{m}}} \leq \widehat{s_{n^{\mathrm{m}}}} \quad \forall n^{\mathrm{m}} \in \mathcal{R}^{\mathrm{m}} \tag{5.42b}$$

$$\text{and equation (5.23f)}$$

This can also be applied to all derived fairness-specific problems.

### 5.4.6 Summary

The previous sections defined the optimization problem that is used for the basic analysis in this thesis. The problem comprises a joint optimization of on-off power level IfCo, RA, and LA. For simplicity and because that has the most significant influence on the result, LA is here modeled as MIMO mode selection, only. The set of MIMO modes contains all viable combinations of numbers of transmit antennas and numbers of spatial streams. The main constrains of the problem are the available time-frequency resources and compute capacity.

The problem is formulated by a combination of linear equations, which allows for efficient solving even with large numbers of constraints. It is realized as a resource model with three layers. On the highest layer, power allocation resources represent every possible combination of transmitting and non-transmitting BSs. On the second layer, at each BS each power allocation resource is split into one scheduling resource per UE. On the third layer, each scheduling resource is split into MIMO resources, one for each MIMO mode. In this model, the sizes of all resources are variable. They are linked to the sizes of the resources on the remaining layers by constraints.

To reduce the complexity of the model, interference is considered only for a limited set of strongest interferers for each UE. This lowers the number of scheduling and MIMO resources.

The model allocates each MIMO resource to one of the power allocation resources. Thus, the channel capacity and compute effort for each MIMO resource can be precalculated. The achieved rate of a UE is calculated by summing over the capacities of the respective MIMO resources, weighting each with the resource's size. The same is applied for the compute effort.

Based on the formulation of the basic problem, section 5.4.4 defined variants for the four fairness schemes *None*, *PF*, *MinR*, and *MaxMin*. Orthogonal to those, section 5.4.5 introduced variants that limit the flexibility of the optimizer. Instead of representing all-encompassing optimization of IfCo, RA, and LA, these are later used to model the separate adaptation of each of these modules.

## 5.5 Inference from Solutions of the Optimization Problem

The previous section 5.4 defined multiple variants of an optimization problem. In this section, these optimization problems are solved and their solutions are evaluated. The findings from these evaluations then serve as the foundation for the design of the proposed system, which is presented in section 5.6.

This section is structured as follows. First, section 5.5.1 summarizes the applied evaluation methodology. The four subsequent sections present different studies. In section 5.5.2, the fairness schemes defined in section 5.4.4 are compared by solving the problem without limiting the processing effort. This serves to discuss the interdependencies of the components of the optimization problem. Section 5.5.3 shows how an optimal system copes with limited processing resources. The problem variants defined in section 5.4.5 are then applied in section 5.5.4 to evaluate which variables have to be adapted to approach the optimal performance. The last study in section 5.5.5 investigates whether an individual adaptation of LA parameters is required for optimal efficiency or a configuration on cell level is sufficient. Finally, section 5.5.6 summarizes the findings from all studies and derives guidelines for the design of the proposed system.

### 5.5.1 Evaluation Methodology

The evaluations presented in this chapter are based on the system model introduced in section 5.3. To simplify the optimization model, the small system model with seven sites and 21 sector cells is used here. The optimization is performed with a fixed number of UEs and full-buffer traffic. Here, the actual number of UEs is a parameter. The first evaluations, which give some insight into the fairness schemes, have been performed for 105, 210, and 420 UEs, which corresponds to 5, 10, and 20 UEs per cell in average. Later evaluations then concentrate on 210 UEs.

Different fairness schemes, which were introduced in section 3.2.2, have been evaluated. However, most of the evaluations are performed for the *PF* problem only, because that is considered to be a good balance between fairness and system throughput. For most of the following evaluations, the available compute resources are a parameter and constraint for the optimization problem.

The optimization model also requires parametrization. Here, the approximation of the logarithm for the *PF* model has been configured with $N_{\text{pieces}} = 3000$ and $u_{\min} = -15$. The scaling factor for the *MaxMin* system has been set to $\epsilon = 1 \times 10^{-3}$. This is assumed to be a good balance. Larger values tend to distort the primary objective, which is to achieve Max-Min Fairness. For smaller values, the optimizer considers the sum throughput component to be irrelevant for the solution. For the *MinR* system, two values for the minimum rate have been evaluated, namely $r^{\min} = 300\,\text{kbit/s}$ and $r^{\min} = 1\,\text{Mbit/s}$.

The evaluations are conducted in $N_{\text{drops}} = 20$ independent drops per parametrization. A *Java* program based on the *IKR SimLib* [IKRSimLib] and *IKR RadioLib* loads a trace file with small-scale channel effects generated in *MATLAB*. It places the UEs randomly, calculates the large-scale effects of the channels, and determines the serving BSs. The best MCS is derived for each combination of MIMO mode and power states of the relevant interferers. The optimization model is configured with the resulting associations of UEs to serving BSs, data rates, and processing efforts. Then a standard solver (*IBM ILOG CPLEX Optimizer* version 12.6.3) is run to solve the problem. For continuous problems, the optimal solution is calculated in all cases. For integer problems, the solver is configured to stop solving when the gap between the best found integer solution and the bound (i. e. the solution of the relaxed problem) falls below 0.5 %. Relevant metrics are extracted from the optimal solution, which are later aggregated for

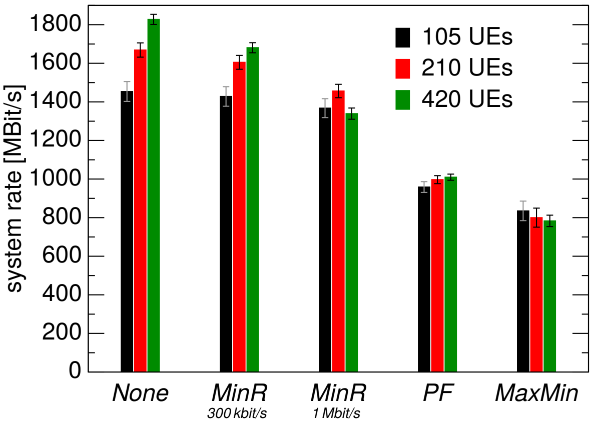


**Figure 5.8:** Total system rate achieved with unrestricted processing effort

evaluation. The placement of UEs and the realizations of the channels are equal for different parametrizations.

The evaluations use the following metrics. The rates achieved by the UEs and derived values of those serve as main metric. First, the average UE rate is used as metric for the general network performance. In general, comparing this is sufficient, because fairness is guaranteed by the formulation of the optimization problems. However, the empirical CDF and the 5th percentile of all UEs' rates are used to get insight into the respective optimal fairness. In addition to the network performance, the occupied processing capacity is also evaluated for those setups where it is not constrained. The objective function of the optimization problem is used at some places to compare different variants of the optimization problem.

For the average UE rate, confidence intervals are calculated with the following procedure. The rates itself are assumed to be not normally distributed. Therefore, first, the average over the rates of all UEs in a drop is calculated. Following the central limit theorem, the resulting drop-averages are approximately normally distributed [Law07, chapter 4.6]. Student's t-distribution is then used to derive the 95 % confidence interval from the drop-averages [Stu08; Law07]. Note that these confidence intervals are approximate due to the limited number of UEs contributing to a drop-average.

### 5.5.2 System Behavior with Unrestricted Processing Effort

The first study is used to get insight into the relations of RA, IfCo, and MIMO mode selection. Especially, it is evaluated how this is influenced by the fairness scheme. In addition, the study is also used to derive the unrestricted compute resource utilization. For this study, the compute resources have not been restricted, i. e. $p_{\max} = \infty$. The optimizer solved the problems specified for different fairness schemes in section 5.4.4 for different numbers of UEs.

Figure 5.8 depicts the sum of the rate of all UEs in the system. Each group of bars represents one fairness scheme, while each color stands for a different number of UEs. Here, the system rate has been chosen as metric to make the number independent of the number of UEs.

As expected, the system rate is highest for the less fair configurations. The highest rate is achieved with the *None* scheme, while the *MaxMin* scheme leads to the lowest rate. Also, with more UEs,

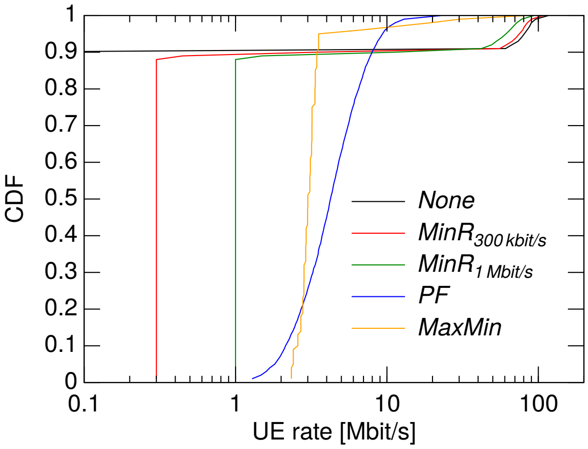


**Figure 5.9:** CDF of UE rates for 210 UEs

the system can make use of the increased diversity of the channel realizations and deliver higher throughput. Only for the *MaxMin* scheme, the rate does not increase with more UEs. For the *MinR* schemes, the effort the system has to invest to guarantee the minimum rate increases with the rate and with the number of UEs. This leads to a reduced system rate for the configuration with $\mathit{MinR}_{\mathit{1\,Mbit/s}}$ and 420 UEs.

Detailed insight into the distribution of the UE rates is provided by figure 5.9 for the configuration with 210 UEs. That figure shows an empirical CDF as introduced on page 55. The range of possible UE rates is plotted on the $x$-axis, the respective fractions of UEs encountered with a smaller rate on the $y$-axis. To be better able to differentiate small rates while also showing high rates, the $x$-axis is scaled logarithmically.

The optimal RA for the *None* scheme assigns zero or very low rates to about 90 % of the UEs. The remaining UEs get all available resources and thereby achieve high rates in the range of 61 Mbit/s to 117 Mbit/s. A similar split into two groups of UEs can be seen for the two *MinR* schemes. There, most of the UEs receive the guaranteed minimum rate, while few remaining UEs maximize their individual throughput with the remaining resources. The throughput of the second group depends on the guaranteed rate, i. e. the higher the guaranteed rate, the lower the higher rates.

The *MaxMin* scheme shows that the maximum possible guaranteed rate in this configuration is about 2.3 Mbit/s. The CDF of this scheme is not a vertical line, because not all UEs' rates depend on each other. While the rate of some UEs cannot be increased further and defines the minimum, the rate of the remaining UEs is maximized by the objective function in problem (5.4.4.4). In addition, the shown CDF is the superposition of all optimized drops, which all have different minimum rates. The smoothest rate distribution is achieved with the *PF* scheme. There, rates are distributed approximately log-normally, with the 5th and 95th percentiles at 1.8 Mbit/s and 9.4 Mbit/s, respectively.

Without compute resource restriction the MIMO modes are always selected to maximize spectral efficiency for the respective channel. At the same time, the optimizer performs IfCo depending on the configured fairness scheme. The resulting allocations are shown in in figure 5.10. Each group of bars results from a different fairness scheme. The three bars in each group stand, from

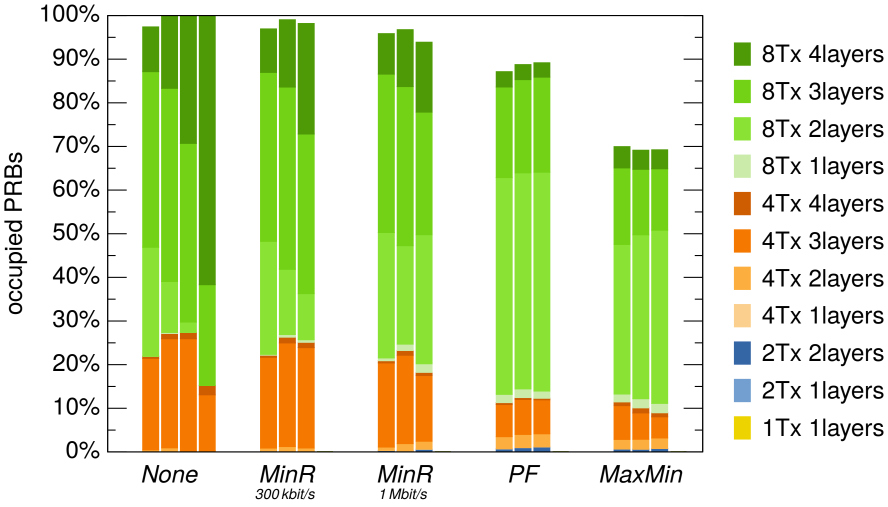


**Figure 5.10:** Utilization of time-frequency resources and MIMO modes

left to right, for 105, 210, and 420 UEs. To give insight into the dependency on the degrees of freedom, a fourth bar is shown for the *None* scheme which stands for 1680 UEs.

The total height of the bars represents the total amount of allocated time-frequency resources. Wherever the bar does not reach 100 %, this is because resources are left free to reduce interference. The heights of the differently colored sections of the bars stand for the amount of time-frequency resources where the respective MIMO mode is used.

With a sufficient number of UEs, the *None* scheme does not perform IfCo but assigns each PRB in each cell to a UE which has favorable channel conditions even with full interference. For the *MinR* schemes, a small amount of time-frequency resources is left free, so that each UE can receive its guaranteed rate. For the fairer configurations *PF* and *MaxMin*, about 10 % or 30 % of the PRBs are idle, respectively.

No configuration uses eight transmit antennas on all PRBs. Instead, 10 % to 27 % of the PRBs are occupied by modes using four virtual transmit antennas, and some even by modes with two antennas. Although counter-intuitive, this effect can be reasoned as follows. Schemes with less than eight transmit antennas are not inferior, but can be interpreted as just a different set of precoding matrices, because UEs have only four antennas. Thus, the preferred number of transmit antennas for a UE only depends on the phase shifts of the respective channel realization. While most channels can be optimally utilized with a precoding from the codebook defined for eight antennas, some benefit from the additional precoding matrices.

The system serves nearly all UEs with two to four spatial streams. This indicates that there is typically sufficient orthogonality between the two polarization planes, while the remaining characteristics of the channel matrix have a higher variance. The most spatial streams are used by the *None* scheme. With increasing number of UEs the probability of experiencing favorable channel conditions increases. Thus, in average, UEs can be served with more spatial streams. The same applies also for the *MinR* schemes, which also assign a part of the resources to those UEs which maximize system throughput. The shares of spatial layers are different for the *PF* and *MaxMin* schemes. There, the largest fraction of resources is used with only two spatial layers. A single spatial stream is only transmitted on about 2 % of the PRBs, and only in those schemes

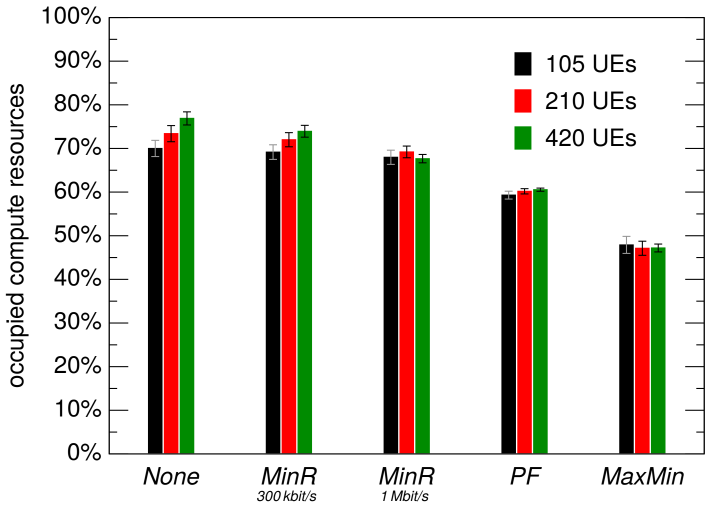


**Figure 5.11:** Occupied compute resources

which have to invest significant resources to provide fairness. In these cases, the optimizer always decides to use eight virtual transmit antennas.

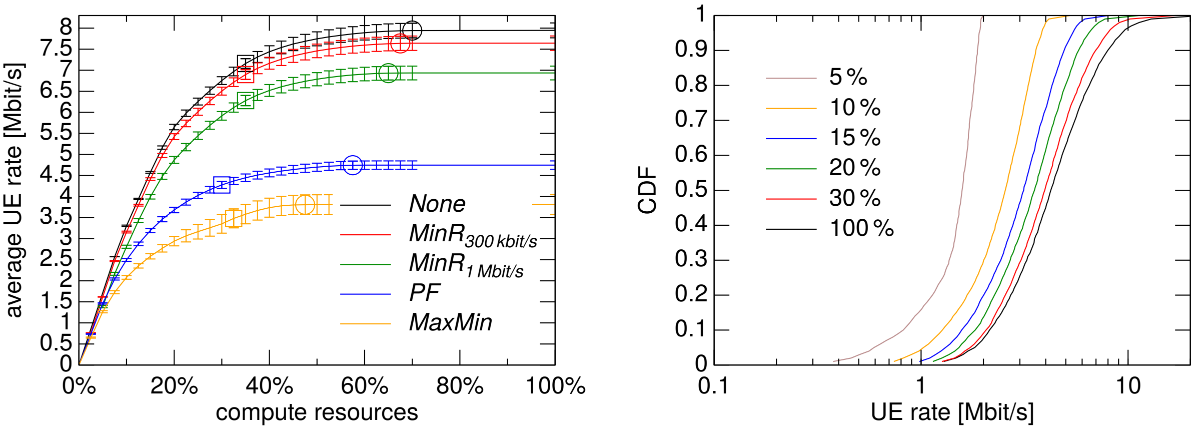
MIMO mode usage and IfCo influence the required processing resources. This is shown in figure 5.11. The configurations occupy 50 % to 80 % of the compute resources.[23] The most resources are required by *None*, while the fairer systems use less resources. The compute resource utilization for the *None* and $\mathit{MinR}_{300\,\mathit{kbit/s}}$ schemes increases with the number of UEs. These resource utilizations are caused by a combination of three effects. First, simpler MIMO modes than the one using eight transmit antennas to transmit four spatial streams cause less effort. Second, the compute effort is reduced whenever a UE is not served with the highest possible MCS. And third, PRBs left free to reduce interference cause no processing effort. It can therefore be expected that the compute resource utilization scales with the aggregated system rate. The closer the system comes to deliver the theoretical peak performance, the more compute resources are required.

The evaluations have shown that the system does not use all compute resources, although the offered load is unlimited. Thus, dimensioning of the compute resources for 100 % compute load is not efficient. Especially for those systems where the operator desires fair serving of the customers, a large amount of compute resources can be saved. Figure 5.11 depicts how many compute resources are required for optimal system operation. However, it does not show whether the system could also operate with less resources. This is studied in the following section.

### 5.5.3 Utility under Restricted Compute Resources

The second study evaluates how the system copes with restricted compute resources. It is evaluated how the system performance degrades when compute resources are limited. In addition, it is evaluated what the main strategies of the system are to manage such a restriction. For this study, the parameter $p_{\max}$ is set to different values to restrict the available compute resources.

[23] The configuration with 1680 UEs and the *None* scheme, for which the PRB utilization is shown in figure 5.10, occupies 87 % of the compute resources.



**(a)** average UE rate

**(b)** CDF of UE rates with *PF* for different compute resource limits

**Figure 5.12:** Evaluation results for optimization problem with restricted processing effort

The optimizer then maximizes the objective function for the respective fairness scheme. To limit the scope of the evaluations, the number of UEs has been fixed to 210.

Figure 5.12a depicts how the average UE rate drops when compute resources are restricted. The available compute resources are plotted on the $x$-axis, the achieved average rate on the $y$-axis. The curves represent different fairness schemes. On each curve, the minimum compute resources which are required to serve the UEs without impact is marked by a circle.[24] The minimum compute resources required to achieve 90 % of the not impacted rate is marked with a square.

In the right third of the plot, more compute resources than required are available. The compute resources required to not impact system performance roughly corresponds to the values shown in figure 5.11.[25] When compute resources become restricted below that value (moving further to the left on the $x$-axis), the system performance is degraded. Here, the evaluated fairness schemes all follow a similar behavior. At first, the average UE rate is reduced only slightly. When the compute resources are limited to values below 40 %, the slopes gradually become steeper. To maintain 90 % of the original rate, the configurations require between 30 % and 35 % of the peak compute resources $P^{\text{peak}}_{|\mathcal{B}|s_{\max}}$.

Compared to the *None* and *MinR* schemes, the *PF* scheme requires less compute resources to achieve full rate and also to maintain 90 % of the full rate. In contrast, the *MaxMin* scheme requires even less resources to achieve its full rate, but is more severely impacted when further limiting the compute resources.

For the *PF* configuration, the performance degradation simultaneously impacts all UEs. This is depicted in figure 5.12b. Here, the empirical CDFs are plotted for different compute resource limits from 5 % to 100 %. When reducing the compute capacity, the curves are shifted to the left, but keep a similar shape.[26] Only for the top 30 % of the rates, the impact is larger. For 5 %

[24] This is here defined as the point with the lowest compute resources where the average UE rate is degraded by less than 0.1 %.

[25] For some configurations, small reductions of the occupied compute resources can be achieved without impacting performance. Therefore, the marked compute resource levels are slightly lower than those in figure 5.11.

[26] The $x$-axis is scaled logarithmically. Thus, a parallel shift is equivalent to scaling all rates with the same factor.

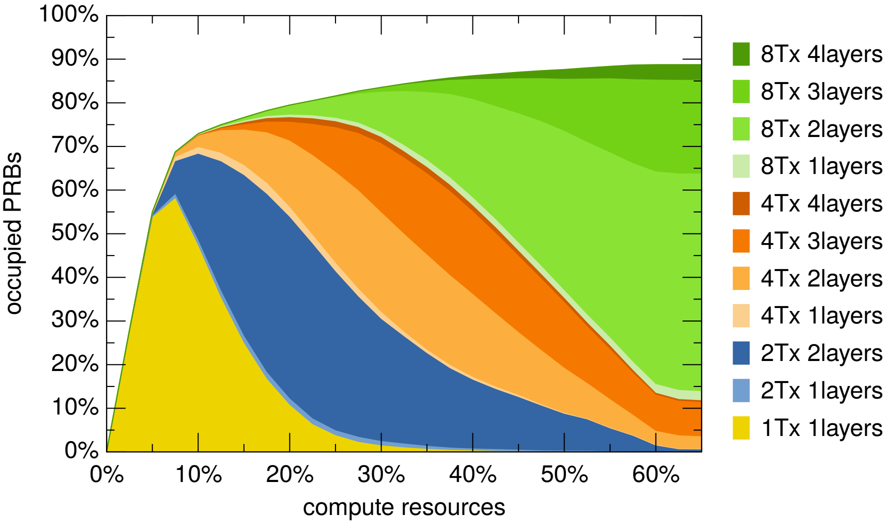


**Figure 5.13:** Utilization of time-frequency resources and MIMO modes with *PF*

compute resources, it seems not efficient for the system to maintain high rates for some UEs. Instead, the CDF becomes steeper, i. e. the rates achieved by the UEs become more similar.

To cope with limited compute capacity, the system adapts IfCo, RA, and MIMO mode utilization. Figure 5.13 shows this for the *PF* configuration. Here, the height of the differently colored areas represent the amount of time-frequency resources where the respective MIMO mode is used. The total height of the colored area is equivalent to the totally used time-frequency resources. Analogously, the height of the white area at the top of the plot stands for the time-frequency resources which are left free and serve to reduce interference.

The fractions plotted for 65 % (on the right side of the plot) are the same as those of the bar for *PF* and 210 UEs in figure 5.10. Thus, there is not change in system behavior when allowing to use more than 65 % compute resources. However, when compute resources are limited to values below that, totally allocated resources are reduced and shares of MIMO modes change. Different behavior of the system can be identified in different parts of the plot.

When reducing the compute capacity from 60 % down to 35 %, the use of MIMO modes using eight transmit antennas is reduced. This is balanced by increasing usage of those modes with four and two transmit antennas. However, these are not significantly used to transmit four spatial layers. Simultaneously to the shift in MIMO mode usage, the fraction of empty resources increases with a low slope.

This slope steepens when further limiting the compute resources down to about 10 %. At the same time, the MIMO mode utilization gradually switches first from four to two antennas, and then also to one antenna. The peak resource usage for four antennas is at 35 %, that for two antennas at 20 %. This comes along with using only two or even one spatial layer. However, it can be seen from the plot that only on few resources a single spatial layer is used except on those where only one transmit antenna is available. Thus, reducing the number of transmit antennas seems more efficient than reducing the number of spatial layers more than necessary. It is also visible that for a single compute resource limit, the system uses a broad range of MIMO modes to serve the different UEs.

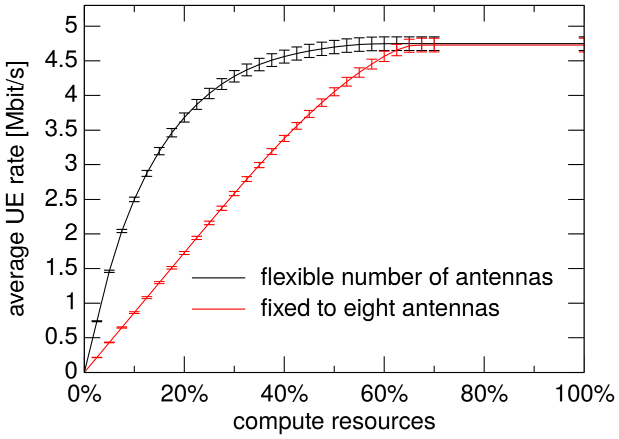


**Figure 5.14:** Impact of not changing the number of virtual transmit antennas

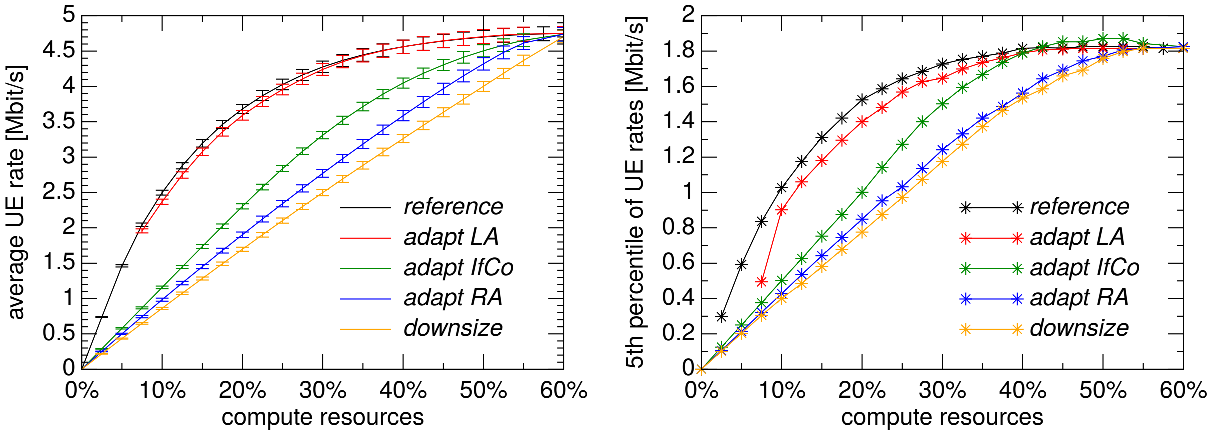
Below 10 %, most PRBs are occupied with the MIMO mode that uses only one transmit antenna. Consequently, the only way to further reduce the processing effort is to leave more PRBs free.

The previous evaluation already indicates that the optimizer heavily relies on the ability to reduce the number of virtual transmit antennas to cope with limited processing capacity. The benefit of this flexibility is also illustrated by figure 5.14. There, the system is optimized for the *PF* scheme in two different configurations. The one with flexible number of virtual transmit antennas is the same as studied previously, i. e. the curve is the same as the respective one in figure 5.12a. For the configuration which is fixed to eight virtual transmit antennas, only those MIMO modes have been allowed to be used by the optimizer which do not reduce the number of transmit antennas.

For compute resources above 70 %, fixing the number of transmit antennas results in a reduction of the average UE rate of about 0.4 %. This is caused by the reduced set of precoding matrices as discussed in section 5.5.2. When compute resources are reduced below 60 %, the network performance of the system which is not allowed to reduce the number of virtual transmit antennas degrades almost linearly. For compute resources below 35 %, the flexible number of transmit antennas gains more than 50 % UE rate compared to the restricted configuration.

Summarizing, the optimizer can maintain relatively high network performance when the compute resources are not limited too severely. When restricting the available compute resources, the performance degrades gracefully. To do so, it simultaneously and gradually reduces the number of virtual transmit antennas and leaves more PRBs free to reduce interference. Not allowing to reduce the number of virtual transmit antennas results in a significant performance degradation. The system uses different MIMO modes at the same time to serve the UEs. Therefore, it does not seem viable to change the number of transmit antennas for the whole system at once. Instead, a more complex strategy seems to be required to achieve the gentle slope in a heuristical implementation.

The evaluations have shown that, when given full flexibility, the optimizer adapts MIMO mode selection as well as IfCo to cope with limited processing capacity. The following section evaluates whether the combination of these strategies is required or it is sufficient to implement only one of them.



**(a)** average UE rate

**(b)** 5th percentiles of UE rate

**Figure 5.15:** Comparison of strategies to cope with reduced processing capacity (UE rates)

### 5.5.4 Separation of Control Variables

Objective of the third study is to evaluate whether the combined adaptation of IfCo, RA, and LA parameters is required to efficiently cope with limited compute resources. As in the previous study, the parameter $p_{\max}$ is set to different values. However, here not the flexible optimization problems from section 5.4.4 are solved, but the restricted versions defined in section 5.4.5. These model that only single modules of the BBU system are adapted to cope with limited compute resources. To simplify the evaluations, the number of UEs is fixed to 210. In addition, only the *PF* configuration is evaluated.

The average UE rate and the 5th percentile of the UE rates are shown in figures 5.15a and 5.15b, respectively. Here, the black curve, denoted as *reference*, resembles the unrestricted optimization problem as evaluated in the previous section.[27] The remaining curves represent the restricted problems from section 5.4.5.

In general, the average rates of the different versions of the optimization problem show a similar behavior. With compute resources above 60 %, all show the same performance (not shown in the plot). The average rates degrade when compute resources are limited further. However, the efficiency with which the variants cope with limited compute resources differs. The performance of *adapt LA*, which adapts only MIMO mode selection, is close to the reference. However, with this strategy it is not possible to reduce the required compute capacity to values below 7.5 %. Thus, the optimization problem is infeasible, and the respective points are missing from the plot. The second highest average rates are realized by *adapt IfCo*. By reassigning resources and by making use of empty resources to reduce interference, a comparatively high performance can be maintained. Similarly, *adapt RA* reassigns time-frequency resources. However, there the optimization problem does not take advantage of empty resources. This results in further degradation of the average rates. As expected, the lowest performance is achieved with the simplest variant *downsize*. There, the average rate depends almost linearly on the available compute capacity.[28]

[27] It is the same as the curve for *PF* in figure 5.12a.

[28] Note that, as not all sets of allocated resources have to be scaled down by the same factor, strictly linear dependence is not necessary.

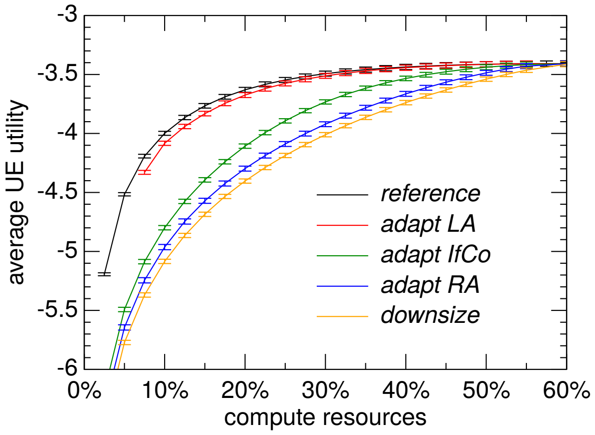


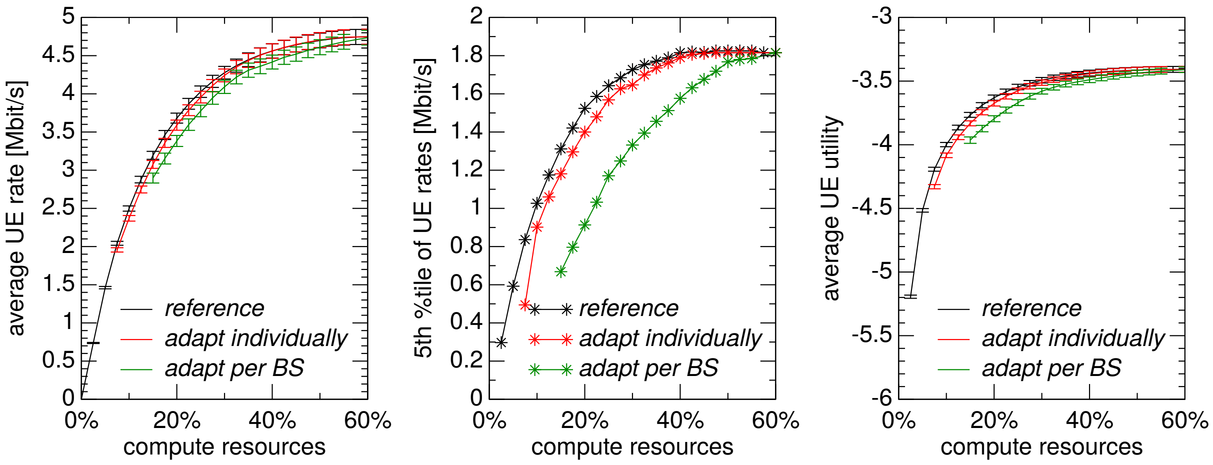
**Figure 5.16:** Comparison of strategies to cope with reduced processing capacity (UE utility)

The percentiles of the UE rates in figure 5.15b show a similar behavior. The most striking difference is that for the curves denoted with *adapt IfCo*, the percentiles of the rates increase slightly when lowering the compute resource limit from 60 % to 50 %. This behavior can be explained by two opposing effects: The variant *adapt IfCo* copes with limited compute resources by transmitting on fewer channel resources. Thus, with fewer available compute resources, the average interference experienced by the system decreases. At the same time, fewer channel resources can be used to serve the UEs, which impacts the experienced rates. In average, the second effect dominates the first one. Thus, in figure 5.15a, the average rate decreases when the compute resources are limited. However, for UEs that suffer heavily from interference, the first effect can dominate the experienced rate. As these UEs do also typically experience low rates, this effect shows up in the percentiles plotted in figure 5.15b. The same behavior also occurs for the *reference*. However, for this variant the difference is barely noticeable from the plot. This behavior does not arise with the other strategies, because those do not reduce interference.

For compute resources below 40 %, the cell border performance of *adapt LA* is lower than that of *reference*, while the average rates are still the same. This indicates that the change of MIMO modes impacts the cell border users, which cannot be compensated by reduced interference or adapted RA in this configuration. Especially for 7.5 % compute resources, this impact is significant. Compared to the average rates, the *adapt RA* performs worse. The percentiles achieved with that strategy are similar to that of the *downsize* variant.

By tuning the fairness configuration, average throughput can be traded for cell-border throughput and vice-versa. A fair comparison of the strategies *adapt LA* and *adapt IfCo* is not possible with these two metrics alone, because each performs better in one of the metrics. Therefore, the PF utility, which represents the objective function of the *PF* configuration, is plotted in figure 5.16.[29] There, it can be seen that the objective achieved with *adapt LA* is significantly better than that of all other strategies. Thus, by shifting resources to cell border users, the 5th percentile performance achieved by *adapt IfCo* can possibly also be achieved by *adapt LA*. However, that would reduce the utility and thereby violate the definition of PF.

[29] Note that the absolute values of the utility are negative. This is caused by the normalization as defined in equation (5.26). In this configuration, there are in average ten UEs served by each cell. Furthermore, as the radio channels are not ideal, the UEs can not be served with highest spectral efficiency. Both effects cause the average experienced rate $r$ to be significantly lower than the maximum rate achievable in a cell $r^{\max}$. This leads to utility values of about -3.4 for the configuration with unconstrained compute resources.



(**a**) average UE rate (**b**) 5th percentiles of UE rate (**c**) Average UE utility

**Figure 5.17:** Benefit of individually adapting LA parameters

Summarizing, in these evaluations the strategy *adapt LA* showed the best performance, which is close to the reference for average rates and utility. A small performance drop is seen for cell border users, however there the performance is still better than that of the other strategies for the largest range of evaluated compute resource limits. The second best strategy is *adapt IfCo*, which favors especially cell border UEs. A significant drawback of the strategy *adapt LA* is that it cannot cope with arbitrary low resource limits. Thus, to achieve stable system operation in all cases, it has to be complemented with a fallback mechanism.

In this section, it was assumed that LA parameters are adapted individually for each UE. The next section evaluates whether this flexibility is required to maintain high network performance.

### 5.5.5 Benefit of Individual Adaptation

The objective of this study is to evaluate whether individual adaptation of LA parameters is required or adaptation on a coarser level is sufficient. Individual adaptation brings fine-grained control and results in high network performance as shown in the previous section. However, it comes with additional complexity as a compute capacity dependent decision has to be made for every UE which has PRBs assigned. When the number of transmit antennas is limited per BS, that simplifies the system. It allows to omit assembly and further processing of the OFDM frame which is to be transmitted on the disabled antennas, given that no other signals (e. g. CSI-RSs) use these antennas.

The evaluations are performed in the same manner as those in the previous study. However, this section focuses on the comparison of the performance achieved by the two optimization problems defined in section 5.4.5.3. The problem which individually adapts LA parameters is identical to the one studied in the previous section.

Figures 5.17a, 5.17b, and 5.17c depict the average UE rates, the 5th percentile of the UE rates and the average PF utility, respectively. The curves *reference* and *adapt individually* correspond to the respective curves in the evaluations in the previous section. The curve *adapt per BS* resembles the optimization problem where the same restrictions are performed simultaneously for all UEs served by a BS. As expected, limiting the flexibility of the optimizer results in degraded network performance. While the degradation of average rates is small, the cell border throughput is impacted significantly. For example, with 40 % compute capacity available, *adapt individually* maintains 99 % of the original rate, while *adapt per BS* achieves only 87 %. This indicates that simultaneously reducing the number of transmit antennas harms especially those UEs which have low channel quality. The average PF utility is also degraded with *adapt per BS*, although the difference is not as large as for the percentile.

Summarizing, individually adapting LA parameters here shows to be beneficial compared to an adaptation per BS. Especially for UEs with unfavorable channel conditions, the simplification comes with a significant impact. However, the advantages of the adaptation per BS cannot be captured by the metrics evaluated here. Thus, in scenarios different to the one considered here, adapting LA per BS can be beneficial.

### 5.5.6 Summary and Derivation of Design Guidelines

Different evaluations of optimization problems have been presented by the previous sections. Section 5.5.2 has shown that, even if the compute resources are not restricted, only 50 % to 80 % of these resources are used. This utilization clearly depends on the fairness scheme, where fairer systems use less processing resources.

Section 5.5.3 evaluated how the system copes with a restriction of the processing resources. It has shown that 30 % to 35 % of the compute resources are sufficient to achieve 90 % of the unrestricted rates. Furthermore, the evaluations indicate that the system simultaneously switches to simpler MIMO modes and leaves PRBs free to cope with this restriction.

The influence of these strategies is separated by the evaluations in section 5.5.4. That section has shown that the performance achieved by adapting only MIMO modes is close to that of the joint optimization. In contrast, modifications of IfCo and RA cannot maintain high rates when the compute resources are restricted.

Finally, the question arises whether it is required to adapt the MIMO mode for each UE individually. Disabling transmit antennas per cell has advantages such as simpler system design and reduced effort in other processing blocks. These advantages are not captured by the metrics evaluated here. Section 5.5.5 has shown that this simplification comes at the cost of reduced cell border throughput.

These outcomes form the foundation for the design of the system proposed in the following section. The proposed system modifies LA, especially the MIMO modes, as that approach has shown promising performance. The evaluations have shown that adapting the number of virtual transmit antennas is crucial for this performance, so the same mechanism is used in the proposed system. The LA is modified individually for each UE, because that is beneficial for the metrics evaluated here. By modifying LA alone, it is not possible to operate with very few compute resources. This mechanism is therefore complemented by a fallback mechanism which allows the system to operate with arbitrary compute resource limitations.

## 5.6 Proposed System

Based on the findings from section 5.5, the proposed system modifies LA to adaptively reduce computational complexity. The system consists of a distributed heuristic to configure LA parameters, a fallback mechanism to guarantee stable system operation under all circumstances, and a prediction mechanism which configures the heuristic to achieve efficient utilization of the available compute resources under different operation conditions. It is is integrated into the components of an eNodeB such that the distributed heuristic acts after IfCo and RA have been performed, but before data is encoded.

Section 5.6.1 states the requirements that the system should meet. The proposed system is derived from a known algorithm to solve an optimization problem. To better understand the concepts of the proposed system, section 5.6.2 introduces this optimization problem and the assocciated solving algorithm. The proposed system itself is then presented in section 5.6.3. That section also illustrates its integration with the components of an eNodeB. Subsequently, section 5.6.4 describes the integration with the remaining LTE system, especially how the proposed system interacts with the UEs. A discussion of the proposed system is provided in section 5.6.5.

### 5.6.1 Requirements and Constraints

The main requirement for the proposed system is to maintain high network performance under restricted compute resources. This should be implementable in a realistic LTE system. Especially, the simplifications conducted for the optimization problem, which allow to use fractions of PRBs to serve a UE simultaneously with different MIMO modes, will not be applied here. Besides that, the following additional objectives apply for the design of the system.

In an unmodified LTE eNodeB, LA is a comparatively simple task, because the eNodeB can follow the proposals from the UE in the CQI reports. Compared to that, the modified LA heuristic should not introduce significant computational or architectural complexity into the system. Especially, the selection of LA parameters should not be performed in tight cooperation between all cells of a BBU pool, but in a distributed fashion.

In section 3.5, it was derived that resource allocation is a complex task. It serves for differentiation between vendors of LTE eNodeBs. Therefore, the used algorithms are not publicly known. To avoid additional complexity and allow evaluation independent of concrete RA algorithms, the interaction of the LA heuristic and the RA should be avoided as far as possible.[30]

In addition, the proposed system shall be robust in different manners. It should be capable of coping, not necessarily efficiently, with arbitrary low compute resources, so that it can maintain at least minimum operation even under extraordinary conditions.[31] It should also be able to handle inexact predictions of compute effort.[32]

[30] The evaluation in section 5.5.4 has shown that the performance loss entailed by this restriction is marginal.

[31] Note that, even if not data is actually transmitted, a minimum amount of compute capacity is necessary to execute the proposed heuristic and transmit control signals.

[32] Exactly predicting the computational effort on a general purpose computer system is often not possible. This is caused by various effects like other tasks running in parallel, the operating system, varying cache hit ratios, and memory latencies. Although guarantees are in principle possible by using a RT operating system and carefully

### 5.6.2 Inspiring Optimization Problem

As introduced in section 5.4.5, it is assumed that IfCo and RA are executed before LA. Here, it is not relevant whether IfCo is performed at each subframe, on coarser intervals, or not at all. However, at each TTI RA decides which UE is served on which set of PRBs. The system then has to select which UE is served with which MIMO mode. The objective for this selection is to maximize system performance while not exceeding the limited processing capacity. It is assumed that each UE is only served with a single MIMO mode and MCS at each TTI. This can be formulated as optimization problem, which is similar to the problem (5.38).

The following section specifies this optimization problem. Subsequently, in section 5.6.2.2 an algorithm is presented which solves the *linear programming* (LP)-relaxation of this problem. This algorithm inspires the design of the proposed system. It can be interpreted as centralized variant of the proposed distributed heuristic.

#### *5.6.2.1 Problem Specification*

The original problem (5.38) is defined as optimizing rates for a snapshot of the system state. Without modification of the formulation this can be applied to optimize the data capacity transmitted in a single subframe.[33] While that problem encompasses all UEs in the system, it can be supposed that in the considered subframe only a subset of UEs in the system get resources assigned by RA. In the following the set $\mathcal{U}$ denotes that subset.

The original problem assumes that IfCo and RA have been performed before the problem is solved. This is there formulated by introducing the parameters $\widehat{s_{n^{\mathrm{s}}}}$ with $n^{\mathrm{s}} \in \mathcal{R}^{\mathrm{s}}$, which denote the fixed amount of allocated resources for each combination of power allocation and UE. The variables for the optimization are then the sizes of the MIMO resources $s_{n^{\mathrm{m}}}$ with $n^{\mathrm{m}} \in \mathcal{R}^{\mathrm{m}}$. The restriction of using only a single MIMO mode could be incorporated into that problem by appropriately constraining the allowed sizes of the MIMO resources. However, this approach does not allow to constrain the MCS, so that the same MCS is used for all resources allocated to a UE. Therefore, a new problem is defined here as follows.

The variables of the new optimization problem are binary flags selecting the applied MIMO modes. These flags are here denoted as $\hat{a}_{u,m}$ with $\hat{a}_{u,m} \in \{0, 1\}$ and $\sum_{m \in \mathcal{M}} \hat{a}_{u,m} = 1 \; \forall u \in \mathcal{U}$. Each flag variable is equal to one if UE $u$ is served with MIMO mode $m$ and equal to zero otherwise. For each UE, each MIMO mode comes with a different data capacity and processing effort.[34] When UE $u$ is served with MIMO mode $m$ on the allocated PRBs, $r_{u,m}$ bits can be encoded. This is calculated according to the description in section 5.3.5, taking into account that a single MCS is used for the set of allocated resources. Analogously, the compute effort required to encode the data is denoted as $p_{u,m}$. This can be derived from equation (5.14).

managing the remaining effects, this is a complex task and can result in reduced efficiency of the system. It is therefore beneficial to have a system which does not strictly require exact prediction of the effort, but can cope with certain variations.

[33] Note that, even if the problem formulation is the same, this does not mean that repetitive optimization per subframe necessarily leads to the same long-term results.

[34] Note that the allocated PRBs are fixed and the MCS is always selected to maximize the capacity for a given allocation and MIMO mode. Thus, data capacity and processing effort depend on the selected MIMO mode, only.

Similar to problem (5.38), the system performance is here modeled as utility $\mathrm{U}_{\mathrm{cap}}(\bullet)$, which is now a function of the data capacity. As only one MIMO mode is used per UE, the utility values corresponding to the MIMO modes can also be precalculated. These are here denoted as $v_{u,m} = \mathrm{U}_{\mathrm{cap}}\left(r_{u,m}\right)$. Based on these definitions, the MIMO mode allocation can be formulated as a *multiple-choice knapsack problem* (MCKP):

$$\max_{\hat{a}_{u,m}, u\in\mathcal{U}, m\in\mathcal{M}} \quad \sum_{u\in\mathcal{U}} \sum_{m\in\mathcal{M}} \hat{a}_{u,m} v_{u,m} \tag{5.43a}$$

$$\text{s. t.} \quad \sum_{u\in\mathcal{U}} \sum_{m\in\mathcal{M}} \hat{a}_{u,m} p_{u,m} \leq P^{\text{peak}}_{|\mathcal{B}|s_{\max}} p_{\max} \tag{5.43b}$$

$$\sum_{m\in\mathcal{M}} \hat{a}_{u,m} = 1 \qquad \forall u \in \mathcal{U} \tag{5.43c}$$

$$\hat{a}_{u,m} \in \{0, 1\} \qquad \forall u \in \mathcal{U}, m \in \mathcal{M} \tag{5.43d}$$

The characteristics of an MCKP are discussed in detail by Kellerer, Pferschy, and Pisinger [KPP04]. They specify upper bounds, list exact solution algorithms, and present solution heuristics. The MCKP is known to be NP-hard. As computational efficiency is one of the requirements for the MIMO mode selection algorithm, an exact solution of the problem is not viable here. Instead, the proposed system is based on the LP-relaxation of the problem.

#### 5.6.2.2 Algorithm to Solve the LP-Relaxation of the Problem

The LP-relaxation is formed by replacing the binary variables $\hat{a}_{u,m}$ in problem (5.43) with non-negative continuous variables $\check{a}_{u,m}$. An efficient approach to solve the relaxed MCKP is presented by Sinha and Zoltners [SZ79] as part of a branch-and-bound solution strategy.[35] This approach is followed here.

First, the problem can be simplified by removing dominated and LP-dominated modes. These are defined as follows. A MIMO mode $m$ of UE $u$ is dominated if there exists another mode $\overline{m}$, such that

$$p_{u,\overline{m}} \leq p_{u,m} \text{ and } v_{u,\overline{m}} \geq v_{u,m}, \tag{5.44}$$

i. e., the mode $\overline{m}$ produces higher utility with less compute effort. A MIMO mode $m$ of UE $u$ is LP-dominated if there exists another two modes $\overline{m}$ and $\widetilde{m}$, such that $p_{u,\overline{m}} < p_{u,m} < p_{u,\widetilde{m}}$, $v_{u,\overline{m}} < v_{u,m} < v_{u,\widetilde{m}}$, and

$$\frac{v_{u,\widetilde{m}} - v_{u,m}}{p_{u,\widetilde{m}} - p_{u,m}} \geq \frac{v_{u,m} - v_{u,\overline{m}}}{p_{u,m} - p_{u,\overline{m}}} \tag{5.45}$$

This means that there is a linear combination of MIMO modes $\overline{m}$ and $\widetilde{m}$ which achieves a higher utility with the same compute effort as caused by mode $m$.

It can be shown that, if mode $m$ is dominated for UE $u$, there exists an optimal solution to problem (5.43) with $\hat{a}_{u,m} = 0$ [SZ79]. This also holds for the LP-relaxed variant of the problem. In addition, for modes which are LP-dominated, there exists an optimal solution to the relaxed

[35] The same algorithm is also reproduced by Kellerer, Pferschy, and Pisinger [KPP04], using a different notation.

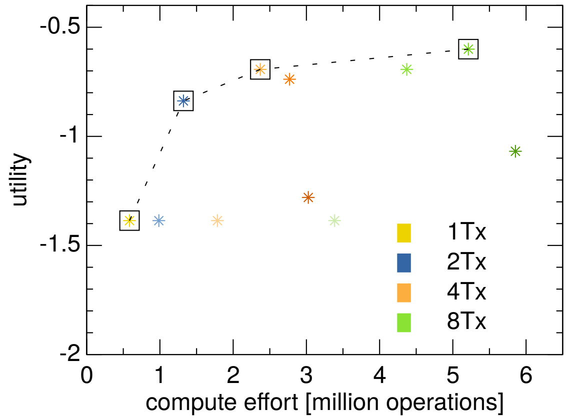


**Figure 5.18:** Exemplary illustration of dominated MIMO modes

variant with $\check{a}_{u,m} = 0$ [SZ79]. Therefore, these modes can be removed from the problem.[36] In the following discussion, the set of those MIMO modes of UE $u$, which are neither dominated nor LP-dominated, is denoted as $\widetilde{\mathcal{M}}_u$.

The concept of LP-dominance is illustrated in figure 5.18. There, the utility is plotted over the compute effort for all modes for a single UE. Colors represent the number of virtual transmit antennas, while shades of the same color denote different numbers of spatial layers.[37] The non-dominated modes are marked with black squares and connected by a dotted line. Note that the dotted line resembles a part of the convex hull of the points. Modes below and right of this dotted line do not need to be considered.

The simplified LP-relaxation can be solved by the following algorithm. First, sort the remaining MIMO modes $\widetilde{\mathcal{M}}_u$ of each UE $u$ by utility (or effort, which is equivalent) in ascending order. Assign indices, such that $r_{u,m_i} < r_{u,m_{i+1}}$ (with $0 \leq i < \left|\widetilde{\mathcal{M}}_u\right|$). For each mode except those with index 0, calculate the additional compute effort as

$$\hat{p}_{u,m_i} = p_{u,m_i} - p_{u,m_{i-1}}. \tag{5.46}$$

In addition, calculate the efficiency of the mode as

$$e_{u,m_i} = \frac{v_{u,m_i} - v_{u,m_{i-1}}}{\hat{p}_{u,m_i}}. \tag{5.47}$$

Note that the efficiency $e_{u,m_i}$ is equivalent to the slope of the dotted line segment leading to the point which represents $m_i$ in figure 5.18.

Start by selecting the mode with least compute effort for each UE, i. e. set $\check{a}_{u,m_0} = 1 \ \forall u \in \mathcal{U}$. Calculate the remaining processing capacity as $p_{\text{rem}} = P^{\text{peak}}_{|\mathcal{B}|s_{\max}} p_{\max} - \sum_{u \in \mathcal{U}} p_{u,m_0}$. In case $p_{\text{rem}} < 0$, the problem (5.43) is infeasible. Otherwise, successively switch to more complex modes. Thereto, first set the remaining candidate modes of UE $u$ to $\widetilde{\mathcal{M}}^{\text{cand}}_u = \widetilde{\mathcal{M}}_u \setminus \{m_0\}$. Then,

[36] Note that for the special case of $p_{u,\overline{m}} = p_{u,m}$ and $v_{u,\overline{m}} = v_{u,m}$, which is included in equation (5.44), only one of the modes can be ignored. As both effort and utility are the same, it is not relevant which of the modes is removed from the problem.

[37] The colors are the same as those in figures 5.10 and 5.13.

from all UEs and all remaining candidate modes, select the combination of UE and mode with the highest efficiency:

$$u_{\text{next}} = \arg\max_{u \in \mathcal{U}} \max_{m \in \widetilde{\mathcal{M}}_u^{\text{cand}}} e_{u,m} \tag{5.48}$$

$$m_{\text{next}} = \arg\max_{m \in \widetilde{\mathcal{M}}_{u_{\text{next}}}^{\text{cand}}} e_{u,m} \tag{5.49}$$

Compare the additional effort for this mode $\hat{p}_{u_{\text{next}},m_{\text{next}}}$ with the remaining processing capacity $p_{\text{rem}}$. In case $p_{\text{rem}} < \hat{p}_{u_{\text{next}},m_{\text{next}}}$, this mode can only be used partially. Set $\check{a}_{u_{\text{next}},m_{\text{next}}} = \frac{p_{\text{rem}}}{\hat{p}_{u_{\text{next}},m_{\text{next}}}}$. Decrease the selection variable for the previously selected mode such that $\sum_{m \in \mathcal{M}} \check{a}_{u_{\text{next}},m} = 1$. The LP-relaxation of the problem is now solved. In case $p_{\text{rem}} \geq \hat{p}_{u_{\text{next}},m_{\text{next}}}$, mode $m_{\text{next}}$ can be used without restriction. Set $\check{a}_{u_{\text{next}},m_{\text{next}}} = 1$ and $\check{a}_{u_{\text{next}},m} = 0 \; \forall m \in \mathcal{M} \setminus \{m_{\text{next}}\}$. Account for the additional effort by setting $p_{\text{rem}} = p_{\text{rem}} - \hat{p}_{u_{\text{next}},m_{\text{next}}}$. Remove the mode from the set of candidates, i. e. set $\widetilde{\mathcal{M}}_u^{\text{cand}} = \widetilde{\mathcal{M}}_u^{\text{cand}} \setminus \{m_{\text{next}}\}$.

Continue by selecting the combination of UE and mode with the highest efficiency from all remaining candidates, until either all UEs use the mode with the highest complexity or the algorithm stops because all compute capacity is occupied. This greedy algorithm is also presented in pseudo-code in algorithm 1.

This algorithm is known to optimally solve the LP-relaxation of problem (5.43) [SZ79]. However, due to the restrictions stated in section 5.6.1, the solution to the LP-relaxation is not directly applicable here. Instead, a possibly non-optimal solution for the non-relaxed problem has to be derived. The simplest approach thereto is to use all modes for which the greedy algorithm has set $\check{a}_{u,m} = 1$. For that UE, for which there are two modes with $0 < \check{a}_{u,m} < 1$, the mode with the lower effort is used. This can be formulated as

$$\hat{a}_{u,m} = \begin{cases} 1 & \text{for } m = \arg\min_{\{\widetilde{m} \in \mathcal{M}:\, \check{a}_{u,\widetilde{m}} > 0\}} p_{u,\widetilde{m}} \\ 0 & \text{otherwise} \end{cases} \qquad \forall u \in \mathcal{U}, m \in \mathcal{M}. \tag{5.50}$$

The resulting total compute effort is always equal to or lower than the available compute capacity $P^{\text{peak}}_{|\mathcal{B}| s_{\max}} p_{\max}$.

Note that this algorithm is not part of the proposed system. However, it can be interpreted as a centralized variant of the system. The following section derives a distributed mechanism from this algorithm.

### 5.6.3 Mechanism to Achieve Elastic Utilization of Compute Resources

The heuristic presented in the previous section, which solves the knapsack problem by deriving a solution from the LP-relaxation, has to be executed centrally for the whole BBU. This contradicts the requirements from section 5.6.1. Thus, the heuristic is converted to a distributed system here.

Section 5.6.3.1 derives the distributed system from the heuristic presented in section 5.6.2.2. The following sections 5.6.3.2 to 5.6.3.4 present the three components of the proposed system in detail. Subsequently, the integration of the components into the remaining system is described in

**Algorithm 1** Solve the LP-relaxation of problem (5.43) following the approaches of Sinha and Zoltners [SZ79] and Kellerer, Pferschy, and Pisinger [KPP04]

```
$p_{\text{rem}} \leftarrow P^{\text{peak}}_{|\mathcal{B}|s_{\max}} p_{\max}$ // initialize: remaining processing capacity set to total capacity
for all $u$ in $\mathcal{U}$ do // initialize: assume mode with least compute effort is used for each UE
    $m_0 \leftarrow \arg\min_{m \in \widetilde{\mathcal{M}}_u} p_{u,m}$ // mode $m_0$ defined to be that with least compute effort
    $\check{a}_{u,m} \leftarrow 0 \quad \forall m \in \mathcal{M}$ // unselect all modes of this UE
    $\check{a}_{u,m_0} \leftarrow 1$ // select mode $m_0$
    $p_{\text{rem}} \leftarrow p_{\text{rem}} - p_{u,m_0}$ // subtract it's effort from the remaining capacity
    $\widetilde{\mathcal{M}}^{\text{cand}}_u = \widetilde{\mathcal{M}}_u \setminus \{m_0\}$ // set of remaining candidates per UE
end for
if $p_{\text{rem}} \geq 0$ then // is the problem feasible?
    done $\leftarrow$ false
    while not done do // successively switch to more complex modes
        $(u_{\text{next}}, m_{\text{next}}) \leftarrow \arg\max_{u \in \mathcal{U}} \arg\max_{m \in \widetilde{\mathcal{M}}^{\text{cand}}_u} e_{u,m}$ // select the combination of UE and mode with the highest efficiency
        if none found then // more complex modes available?
            done $\leftarrow$ true // problem trivial, highest throughput modes used for all UEs
        else if $p_{\text{rem}} \geq \hat{p}_{u_{\text{next}},m_{\text{next}}}$ then // can mode $m_{\text{next}}$ be used without restriction?
            $\check{a}_{u_{\text{next}},m} \leftarrow 0 \quad \forall m \in \mathcal{M}$ // unselect the previously selected mode for this UE
            $\check{a}_{u_{\text{next}},m_{\text{next}}} \leftarrow 1$ // select mode $m_{\text{next}}$
            $p_{\text{rem}} \leftarrow p_{\text{rem}} - \hat{p}_{u_{\text{next}},m_{\text{next}}}$ // subtract the additional effort
            $\widetilde{\mathcal{M}}^{\text{cand}}_u \leftarrow \widetilde{\mathcal{M}}^{\text{cand}}_u \setminus \{m_{\text{next}}\}$ // remove $m_{\text{next}}$ from the set of candidates for this UE
        else // $m_{\text{next}}$ can only be used partially
            $\check{a}_{u_{\text{next}},m} \leftarrow \check{a}_{u_{\text{next}},m} \left(1 - \frac{p_{\text{rem}}}{\hat{p}_{u_{\text{next}},m_{\text{next}}}}\right) \quad \forall m \in \mathcal{M}$ // scale down the previously selected mode for this UE
            $\check{a}_{u_{\text{next}},m_{\text{next}}} \leftarrow \frac{p_{\text{rem}}}{\hat{p}_{u_{\text{next}},m_{\text{next}}}}$ // partially select $m_{\text{next}}$
            done $\leftarrow$ true // processing capacity exhausted
        end if
    end while
end if
return done // if true problem is feasible and solved by variables $\check{a}_{u,m}$
```

section 5.6.3.5. Finally, section 5.6.3.6 discusses the alignment of the strategies of the proposed system and of the RA component.

#### *5.6.3.1 Derivation of the Distributed System*

Algorithm 1, which is described textually on pages 135 to 136, can be reinterpreted as determination of an efficiency threshold. Each UE starts with the most efficient MIMO mode. The algorithm then successively switches to MIMO modes with lower efficiency and higher data capacity. During this process, all modes with an efficiency above a certain value are selected at least temporarily. When it completes, out of these modes those remain selected which provide the highest data capacity for the respective UE.

The selection of MIMO modes can be derived from this threshold as follows. Define the threshold $e_{\min}$ to be the efficiency of the last mode which can be used without restriction.[38] To simplify the formulation, assume that modes with index 0 have infinite efficiency, i. e. $e_{u,m_0} = \infty\ \forall u \in \mathcal{U}$. With this, equation (5.50) can be reformulated as

$$\hat{a}_{u,m_i} = \begin{cases} 1 & \text{for } e_{u,m_i} \geq e_{\min} \wedge \left(i = \left|\widetilde{\mathcal{M}}_u\right| - 1 \vee e_{u,m_{i+1}} < e_{\min}\right) \\ 0 & \text{otherwise} \end{cases} \quad \forall u \in \mathcal{U}, m \in \mathcal{M}. \tag{5.51}$$

This means that a mode is selected if it is allowed by the threshold and either it is the most complex mode of the respective UE or the next complex mode of the respective UE is disallowed. Phrased differently, for each UE the most complex mode is selected out of those modes which have higher efficiency than the threshold. Given that the threshold is known, the steps of removing dominated MIMO modes, sorting the remaining modes, and selecting modes following equation (5.51) can be performed independently for each UE. This forms the foundation for the proposed system.

The value of the threshold depends on the load of the system. In case the system load is low and compute resources are plentiful, a low threshold is used to select computationally demanding MIMO modes. When the system load is higher and / or compute resources are scarce, a higher threshold is required, to restrict the system to use only those MIMO modes which are efficient w. r. t. computational effort. Previous evaluations have indicated that the aggregated load of a cluster of BBUs becomes smooth when the cluster is sufficiently large [WGP13]. It is therefore here assumed that the optimal threshold is similar for consecutive subframes. This suggests that a prediction of the threshold is possible.

The distributed algorithm consists of three components. First, the value of $e_{\min}$ is predicted based on the previous subframes. Second, a selection algorithm is executed independently for each UE, which takes the CSI of the UE and the predicted value of $e_{\min}$ and selects a MIMO mode accordingly. The prediction of $e_{\min}$ may be suboptimal. Therefore, the total compute effort resulting from the selected MIMO modes can exceed the compute capacity. Thus, the third component cancels allocations of PRBs to UEs to bring down the total compute effort below the available capacity. These three components are presented in detail in the following sections.

[38] This means it is the last mode for which algorithm 1 takes the branch of line 16 and the following lines.

#### *5.6.3.2 MIMO Mode Selection*

MIMO modes are selected independently for each UE by the following mechanism. Assume that the value of the threshold $e_{\mathrm{min}}$ has been predicted. Further suppose that CSI is available for the set of resources allocated by RA. Determine the MCS which maximizes spectral efficiency for each MIMO mode. For each MIMO mode, calculate the utility value $v_{u,m}$ and the compute effort $p_{u,m}$. Sort all modes by increasing compute effort and assign indices correspondingly.

Removing dominated and LP-dominated modes and determining that mode which provides the best throughput while having an efficiency higher than the threshold can be combined into a single algorithm. This is described in pseudo-code in algorithm 2.

**Algorithm 2** Select MIMO mode for UE $u$

```
1 $i \leftarrow 0$
2 for $j = 1$ to $|\mathcal{M}| - 1$ do
3     $e_{u,m_j} \leftarrow (v_{u,m_j} - v_{u,m_i}) / (p_{u,m_j} - p_{u,m_i})$
4     if $e_{u,m_j} \geq e_{\mathrm{min}}$ then
5         $i \leftarrow j$
6     end if
7 end for
8 return $m_i$
```

The algorithm repeatedly tests whether to switch from an already accepted mode $m_i$ to a more complex mode $m_j$. The ordering of the modes ensures that the processing effort of $m_j$ is larger than or equal to that of $m_i$. First, assume that modes $m_i$ and $m_j$ are both neither dominated nor LP-dominated. In line 3, the efficiency $e_{u,m_j}$ is calculated. If that exceeds the threshold $e_{\mathrm{min}}$, the index $i$ is updated in line 5. This makes $m_j$ the accepted mode and the base for further efficiency calculations. If it does not exceed the threshold, index $i$ is not updated. Although the algorithm could abort here, evaluating the efficiency of the remaining modes does no harm, because their efficiency cannot exceed $e_{u,m_j}$.

Now, consider the case that $m_j$ is LP-dominated. Here, two cases can be differentiated. Either, the efficiency $e_{u,m_j}$ is below the threshold $e_{\mathrm{min}}$, or it is above or equal. In case $e_{u,m_j} < e_{\mathrm{min}}$, index $i$ is not updated, and $m_i$ is compared to the next mode. Otherwise, i. e. in case $e_{u,m_j} \geq e_{\mathrm{min}}$, mode $m_j$ is considered as the current best mode and index $i$ is updated. However, the efficiency of the following non-dominated and non-LP-dominated mode is larger than $e_{u,m_j}$. Therefore, the algorithm will definitely switch to that mode. So, it is not relevant that mode $m_j$ was considered as best mode in the meantime. Finally, if $m_j$ is dominated, the slope calculated in line 3 is negative, and thus index $i$ is not updated.

Algorithm 2 returns a single MIMO mode. This is used in combination with the corresponding MCS to transmit the data to UE $u$.

#### *5.6.3.3 Encoding Data and Canceling Transmissions*

After the MIMO modes have been selected for the UEs with resource allocations, their data can be encoded. A certain amount of time is provided for the encoding, after which the encoded data has to be forwarded to be assembled to a subframe, be further processed, and finally transmitted.

Denote the effort to encode the data for a UE $u$ as $p_u^{\mathrm{real}}$. This effort depends on the number of allocated PRBs, the channel quality, and the MIMO mode selected by the component described in section 5.6.3.2.[39] The total accruing effort is then $p_{\mathrm{total}}^{\mathrm{real}} = \sum_{u \in \mathcal{U}} p_u^{\mathrm{real}}$.[40] In case $p_{\mathrm{total}}^{\mathrm{real}} > P_{|\mathcal{B}| s_{\mathrm{max}}}^{\mathrm{peak}} p_{\mathrm{max}}$, the system is overloaded despite the adapted MIMO mode selection.

The proposed system handles overload situations as follows. Encoding starts and proceeds in random order[41] until the time provided for encoding is over. All remaining UEs, for which no data could be encoded, get their resource allocations canceled.[42] In case the system is currently encoding data when it runs out of time, the calculations are aborted and the respective UEs also receive no data in the current subframe.

#### *5.6.3.4 Predicting the Threshold* $e_{\mathrm{min}}$

The efficiency of the proposed mechanism mainly depends on the apt choice of the threshold $e_{\mathrm{min}}$. If the threshold is too high, only simple MIMO modes will be used. Compute resources stay unemployed, and network performance is suboptimal. However, if the threshold is too low, overload is caused and many resource allocations have to be canceled. Time-frequency resources are left free, which also impacts network performance.

In this dissertation, the focus does not lie on elaborate prediction methods for the threshold $e_{\mathrm{min}}$. Instead, a simple control loop is applied to determine the threshold based on the compute effort which occurred in previous subframes. After each subframe, the value of $e_{\mathrm{min}}$ is adapted. To avoid frequent skipping due to overload, an offset parameter $p_{\mathrm{off}}$ is introduced, such that $e_{\mathrm{min}}$ can be increased before the compute capacity is exceeded. Thus, when the actual compute effort rises above the reference value, i. e. $p_{\mathrm{total}}^{\mathrm{real}} > P_{|\mathcal{B}| s_{\mathrm{max}}}^{\mathrm{peak}} p_{\mathrm{max}} - p_{\mathrm{off}}$, the system increases the value of $e_{\mathrm{min}}$ by $e_{\mathrm{step}}$, otherwise it decreases it by $e_{\mathrm{step}}$. The parameters $p_{\mathrm{off}}$ and $e_{\mathrm{step}}$ allow to tune the system.

#### *5.6.3.5 Integration into the BBU*

The three components of the proposed system are integrated into the remaining modules of the BBU as depicted in figure 5.19. Black text in the figure denotes functions not adapted to cope with reduced compute capacity. In contrast, red text marks additions and modifications.

Here, steps A and B denote IfCo and RA, respectively. These are not modified. In step C, the heuristic defined in section 5.6.3.2 is used to select a MIMO mode and a MCS for each UE to which the RA has assigned a set of PRBs. Following that, in step D the data to be transmitted to the UEs is encoded as described in section 5.6.3.3. Step E controls that the previous two steps are repeated until the system runs out of compute time. If overload occurs, steps C and D are skipped for the remaining UEs. Subsequently, in step F the radio frame is assembled and further calculations are performed. Finally, in step G the subframe is transmitted. After each

[39] Ideally, $p_u^{\mathrm{real}}$ is equal to the compute effort assigned to the selected mode in section 5.6.3.2. However, deviations could be caused by inaccuracies in the compute effort model or by unpredictable overhead.

[40] Note that, as defined in section 5.6.2.1, $\mathcal{U}$ comprises only those UEs which got resources assigned by RA for the current subframe.

[41] Random order has been chosen here for simplicity. See also the discussion in section 5.6.5.3.

[42] For a discussion of the impact on the signaling between BS and UEs, see section 5.6.4.3.

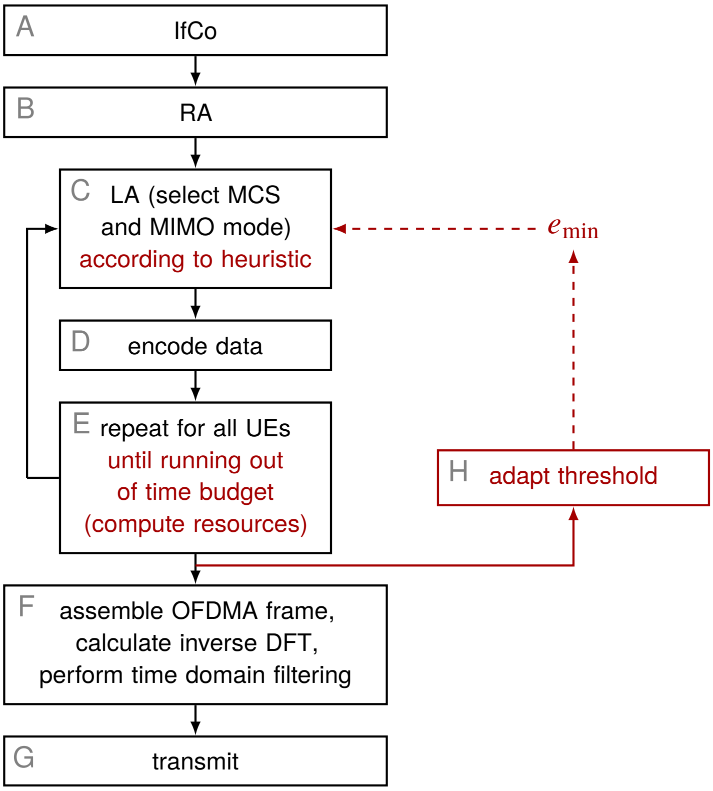

A IfCo
B RA
C LA (select MCS and MIMO mode) according to heuristic
$e_{\min}$
D encode data
E repeat for all UEs until running out of time budget (compute resources)
H adapt threshold
F assemble OFDMA frame, calculate inverse DFT, perform time domain filtering
G transmit

**Figure 5.19:** Schematic picture of distributed heuristic integrated into the BBU

subframe has been processed, in step H the variable $e_{\min}$ is adapted according to the overload which occurred in step D as stated in section 5.6.3.4.

### *5.6.3.6 Alignment with Resource Allocation Policies*

Although the RA module is not modified to cope with limited compute capacity, the proposed system indirectly interacts with RA. The selection of MIMO modes decides the apportionment of compute resources between multiple UEs. This influences the fairness of the system, analog to the allocation of time-frequency resources. Contradicting actions of RA and MIMO mode selection could impact the network performance.

For example, RA could choose to assign a large number of PRBs to a UE with low channel quality. For the same UE, the LA module might select a MIMO mode with low complexity and low spectral efficiency. The saved compute resources could be used to serve other UEs and thereby maximize total data capacity. Thus, the result would be a significant amount of time-frequency resources invested, but missing compute resources render their usage inefficient.

To avoid such contradiction, the utility function used in the proposed system has to be aligned with the RA policy. The utility $\mathrm{U}_{\mathrm{cap}}(\bullet)$ introduced in problem (5.43) is similar to the utility used in the resource allocation optimization problems discussed in section 3.3. However, here the utility function is applied on data capacities per subframe, while the previous one is a function of a long-term data rate. Nevertheless, the logarithm is used here to achieve a fairness of compute resource allocation which matches the PF heuristic used to allocate time-frequency resources. Thus, here

$$\mathrm{U}_{\mathrm{cap}}(r) := \log\left(\frac{r}{r^{\max}}\right), \tag{5.52}$$

where $r$ is the data capacity of the allocated PRBs with a certain MIMO mode and $r^{\max}$ the maximum capacity per subframe. This definition would have to be adapted if a different RA policy than PF is applied.

### 5.6.4 Integration into the RAN

The proposed system does not interact only with the modules of the BBU, but also with other components of the LTE system. The impact on the EPC is low, because varying data rates which are under control of the eNodeBs are a core concept of LTE. In contrast, on the air interface the proposed system behaves differently as an unmodified eNodeB. Aspects of this interaction with UEs are discussed in the following sections.

#### *5.6.4.1 Signaling the Number of Transmit Antennas*

The proposed system relies on a flexible selection of MIMO modes to reduce computational load. To decode the received data, UEs have to know how it was transmitted. As introduced in section 2.3.4.4, MIMO operation in LTE is based on the definition of TMs. Each UE gets a single TM assigned by a semi-static configuration. Depending on the configured TM, different parameters of the encoding can be changed dynamically.

TM 4 was introduced in LTE Release 8 for closed-loop codebook based precoding with up to four transmit antenna ports.[43] It allows the eNodeB to configure the number of spatial layers, and also to fall back to SFBC. However, it does not allow to configure the number of transmit antennas. Instead, that is configured statically for the whole cell. In Release 10, TM 9 was introduced, which allows to use spatial multiplexing with up to eight transmit antenna ports. It is based on DM-RSs for demodulation, so that the applied precoding does not need to be communicated to the decoding UE. This TM allows the eNodeB to apply an arbitrary precoding matrix. It can thus be used to flexibly change the number of transmit antennas. The same applies for TM 10, which was introduced in Release 11.

Compared to TM 4, the DM-RSs used in TMs 9 and 10 come with a certain overhead, which can impact network performance. However, this overhead has to be accepted to be able to use up to eight transmit antennas. These TMs are not applicable for legacy UEs. For those, which have to use TM 4 for closed loop spatial multiplexing, the system can only switch to SFBC to reduce computational complexity. However, as the major load is assumed to be caused by the more complex transmissions used to serve newer UEs, this is not studied in this thesis.

[43] See table 2.4 for an overview of all TMs.

#### *5.6.4.2 Acquiring Channel State Information*

The proposed system relies on the fact that CSI is available for different MIMO modes (see section 5.6.3.2). As described in section 2.3.6.3, CSI measurement can either be based on CRSs or on CSI-RSs. To use up to eight transmit antennas, CSI-RSs have to be applied. UEs measure the channel and report the measurement results by transmitting a RI, a PMI and one or two CQIs. These reports refer to the antennas which transmit CSI-RSs and to the number of spatial layers encoded in the RI.

To get reports for different hypotheses of transmit antennas, TM 10 allows to configure multiple CSI processes. Thus, the eNodeB can transmit different sets of CSI-RSs with different numbers of virtual transmit antennas. UEs can be configured to measure these separately and provide independent CSI reports.

The number of spatial layers which results in maximal spectral efficiency for a certain CSI process is determined by the UE. However, this is only meant as recommendation for the eNodeB. The eNodeB is also allowed to transmit a lower number of spatial layers, thereby overriding the recommendation of the UE.[44]

Due to the special design of the precoding matrices, the reported PMI is also applicable for a lower number of spatial layers. However, the reported CQIs cannot directly be applied to the reduced number of spatial layers. As each spatial layer can provide different performance, omitting a single layer may either increase or decrease the total channel capacity. Omitting a layer decreases interference between layers and can thereby improve the channel quality experienced on the remaining layers. As described in section 2.3.3.1, the eNodeB can apply an offset to the reported CQI to achieve a desired average decode probability. This is meant to compensate for UE-specific deviations, but can also be used to adapt the CQI to a reduced number of spatial layers. The involved impact on performance, which results from inaccurate estimation of the channel capacity, is not considered in this thesis.

Transmitting multiple sets of CSI-RSs with different numbers of transmit antennas causes overhead and thereby impacts network performance. This can be minimized by extending the transmission intervals, however that comes with reduced measurement accuracy. In case of a rather static utilization of the compute capacity, some configurations of transmit antennas can potentially be temporarily disabled. When these configurations then shall be used again, changes to the semi-static configuration of the UEs are required. A similar concept applies for CSI reports for multiple CSI processes. There, the eNodeB can selectively request reports for different CSI processes (see also section 2.3.6.4). It can thereby dynamically trade overhead for more elaborate channel information.

#### *5.6.4.3 Canceling Allocated Resources*

In case the threshold $e_{\text{min}}$ is predicted too low, the proposed system selects too complex MIMO modes. To cope with this situation, UEs which got sets of PRBs assigned by the RA module get these allocations canceled. For higher layers, this cannot be differentiated from the case that a

[44] This is sometimes termed *rank override* [DPS16].

transmission could not be decoded or did not take place at all. However, the interactions with the lower layers of the air interface have to be discussed. There, the canceling of allocated resources can be realized in two different ways, depending on when the PDCCH is encoded.

The PDCCH can be encoded when the OFDMA frame is assembled (i. e. as part of step F in figure 5.19). In this case, the DCIs of a canceled allocation can be omitted from the PDCCH (see section 2.4.3.2). Thus, the intended receivers do not recognize the allocation, but further await transmissions in the following subframes. It is irrelevant what is transmitted on the respective symbols in the PDSCH, because these are not decoded by any UE. The RA module in the eNodeB should recognize that the transmission was skipped, so that it can allocate new PRBs in a following subframe.

Encoding the PDCCH could also be implemented to start immediately after RA (i. e. after step B in figure 5.19).[45] In this case, it is not be possible to remove the DCIs when an allocation is canceled. Consequently, one or multiple DCIs are transmitted for which the corresponding data is missing from the PDSCH. The receiving UEs will recognize this as decode errors. The respective HARQ process at the UE now holds useless data in its buffer, which should not be combined with the next transmission. The eNodeB can circumvent that by instructing the UE to discard the previously received data.[46] As the receiving UE does not get any other information from the received DCI, the remaining protocols are not disturbed.

When resource allocations are canceled, it does not make sense to transmit any PDSCH data symbols on the respective PRBs. However, at this point in the processing, it is too late to make use of this fact by deliberately adapting the MCS in neighboring cells. Thus, only the statistically lower interference can be used. In the aspired operating point of the proposed system, where the fraction of PRBs not used because of compute resource overload is low, this can only bring marginal gain.

### 5.6.5 Discussion and Evaluation

Section 5.6.1 stated a number of requirements for the designed system. The focus of this section is to qualitatively assess the proposed system w. r. t. these requirements. The complementing quantitative evaluation of the network performance is provided in chapter 6. The requirements can be divided into the topics network performance, implementation complexity, interaction with RA, and robustness. They are covered by the following sections in this order.

#### *5.6.5.1 Performance*

The main objective of the proposed system is to maintain high network performance even with limited compute capacity. By design, the proposed system does not impact the performance of the network in case no effective compute resources limitation occurs. However, in case the resources are restricted, the proposed system cannot be expected to achieve the same performance

[45] This can be beneficial, because the PDCCH can then be processed in parallel to the UE-related operations (steps C to E). This also allows to calculate the inverse DFT for the first OFDM symbols of a subframe before the OFDMA frame is assembled. Thereby, utilization of compute resources may be improved.

[46] For details about HARQ operation see section 2.3.3.2.

as the combination of optimal IfCo, RA, and LA. Four different causes for reduced performance are presented by the following paragraphs.

First, the proposed system operates consecutively for every subframe. It can therefore not anticipate future variations of external conditions (e.g. channel variations). Furthermore, the system applies its objective to the individual subframes. In principle, it is possible to evaluate convergence behavior theoretically [Sto05]. However, that is out of scope of this thesis. Instead, this is here evaluated empirically in section 6.1.3.

Second, the proposed system adapts only MIMO modes, while leaving IfCo and RA unmodified. The impact of this restriction has already been evaluated in section 5.5.4. It is picked up again in the evaluation in section 6.1.3.

Third, reduced performance is expected to be caused by the fact that the proposed system resembles a greedy heuristic solving the MCKP. In the general case, this greedy heuristic has arbitrary bad worst-case performance.[47] However, performance is better if the effort for a single UE is low compared to the total compute capacity: The LP-relaxation of the MCKP discussed in section 5.6.2.2 provides an upper bound of the achievable utility. Relative to that, the greedy solution derived in equation (5.50) loses the utility gained by the partially selected mode of one UE. The additional compute capacity used by this mode is left unused, while the remaining capacity is invested efficiently.[48] The amount of unused compute capacity can be as large as the additional effort required for a MIMO mode for a single UE. In contrast, the optimal solution can possibly use this capacity efficiently. The maximum difference in terms of utility between the greedy algorithm and the upper bound therefore is

$$v_{\text{lost,max}} = \left( \max_{u \in \mathcal{U}, m \in \mathcal{M}} \hat{p}_{u,m} \right) \left( \max_{u \in \mathcal{U}, m \in \mathcal{M}} e_{u,m} \right), \tag{5.53}$$

with $\hat{p}_{u,m}$ and $e_{u,m}$ calculated as in equations (5.46) and (5.47), respectively. Given that the threshold $e_{\min}$ is defined optimally, the proposed system achieves the same performance.

The fourth cause for performance impacts is the prediction of the optimal efficiency threshold $e_{\min}$ and the associated canceling of resource allocations in case overload is caused by a too low threshold. In general, there is a trade-off between low utilization of compute resources and the canceling of transmissions. A rather lower value of $e_{\min}$ results in more complex MIMO modes. Thus, the utilization of the compute resources is high, but overload occurs frequently. In contrast, less complex MIMO modes (higher $e_{\min}$) mostly avoid canceling of transmissions at the cost of underutilized compute resources. Possible errors in the prediction of the efficiency threshold $e_{\min}$ can result from the following three causes.

First, assume constant compute load, i. e. channels and RA do not change. The mechanism to predict $e_{\min}$ was proposed in section 5.6.3.4. Assuming that the offset is not applied, i. e. $p_{\text{off}} = 0$,

[47] Kellerer, Pferschy, and Pisinger [KPP04] provide an example, which is similar to the following: Assume two UEs with two MIMO modes each, and a total compute capacity of $P^{\text{peak}}_{|\mathcal{B}|s_{\max}} p_{\max} = x$, with $x > 2$. For both UEs, $m_0$ delivers zero utility with zero effort. For UE $u_0$, $m_1$ delivers utility 2 with effort 1, while for UE $u_1$, $m_1$ delivers utility $x$ with effort $x$. The greedy heuristic will select $m_1$ for UE $u_0$ and $m_0$ for UE $u_1$. This results in a total utility of 2, while the optimal solution has utility $x$.

[48] Compare to the example provided in note [47]: There, the unused compute capacity is $x - 1$. In contrast to the greedy solution, the optimal solution can make use of this capacity to increase the total utility by $x - 2$.

the total accruing compute effort $p_{\text{total}}^{\text{real}}$ oscillates around the total compute capacity $P_{|\mathcal{B}|s_{\max}}^{\text{peak}} p_{\max}$. Thus, marginal overload occurs every second subframe. Whenever that happens, the encoding for at least one UE is skipped. The restricted granularity of this mechanism causes some compute capacity to remain unused. However, that does not imply a significant performance drop as long as the processing effort to encode a single UE is low compared to the totally available compute capacity. In addition, the frequency of these overload situations can be reduced by configuring an offset $p_{\text{off}} > 0$, at the cost of reduced resource utilization.

Second, assume that the load of the network is constant, but variations of the compute load are implicated by small-scale channel fading and actions of the RA algorithm. These fluctuations cannot be anticipated by the proposed prediction mechanism. The impact of this effect is evaluated in a scenario with constant network load in section 6.1.

Finally, when not even the load of the network is constant, the magnitude of the fluctuations increases. The performance of the proposed system then depends on the capability of the control loop to dynamically adapt the value of $e_{\min}$ to the system conditions. A thorough control-theoretic evaluation of the control loop is out of scope of this thesis. Instead, the performance of the complete system is evaluated in a scenario with dynamic traffic in section 6.2.

Summarizing, it is expected that the proposed system does not achieve optimal performance due to its heuristic nature. A quantitative performance evaluation is provided in chapter 6.

#### *5.6.5.2 Complexity*

The proposed system is used to reduce the computational effort required to operate a set of LTE cells. However, the system itself also introduces additional complexity, which counteracts this goal. Thus, the introduced complexity should not exceed the savings. In addition, it should be predictable or at least manageable to facilitate efficient dimensioning of compute resources.

The proposed system consists of three main components. The following paragraphs focus on the MIMO mode selection described in section 5.6.3.2. The remaining two components are the mechanism to cancel the encoding when running out of compute capacity (section 5.6.3.3) and the prediction of the threshold (section 5.6.3.4). Both do not contribute significant effort.

The newly introduced algorithm 2 is executed once per subframe for each UE which got resources allocated by RA. As it iterates over the list of all MIMO modes, it's total effort scales linearly with the number of UEs with resources $|\mathcal{U}|$ and with the number of MIMO modes $|\mathcal{M}|$.

In addition, a precondition for the algorithm is that for each UE, the MIMO modes are sorted by compute effort. This scales with $|\mathcal{U}| \cdot |\mathcal{M}| \log |\mathcal{M}|$, which dominates the complexity of algorithm 2. However, ordering with random input sequence is not required. This is caused by the fact that MIMO modes typically have a fixed order w. r. t. compute effort, which is determined by the number of spatial layers and the number of transmit antennas. Only in rare cases, the influence of the MCS on the compute effort results in a different order. Therefore, static pre-sorting of the modes practically reduces the effort for sorting to that for a simple linear check for the correct order, i. e. the effort now scales with $|\mathcal{U}| \cdot |\mathcal{M}|$.

In a real system, both $|\mathcal{U}|$ and $|\mathcal{M}|$ are limited to small values. The number of MIMO modes is limited by system design.[49] While more modes in general provide finer increments in effort and capacity, they also come with overhead for CSI acquirement. The number of UEs which get resources assigned are constrained by the capacity of the PDCCH as discussed in section 2.4.4. It is thus assumed that the complexity for MIMO selection, i. e. for sorting of MIMO modes and for execution of algorithm 2, can be neglected.

Besides the MIMO mode selection algorithm, additional effort occurs to gather the required input data. To calculate the capacity of the allocated PRBs, the MCS has to be derived from the UEs CSI reports. To allow the proposed system to select from different MIMO modes, this has to be performed for each mode.

There are several ways to further reduce this effort. First, based on the CSI reports of the UE, some modes can be identified to be not efficient. Assume that a UE reports a RI for each number of virtual transmit antennas. All modes where the number of spatial layers exceed the respective RI can be assumed to be dominated, and thus be ignored for the further processing. Second, a slowly operating control mechanism could be used to limit the number of candidate modes. That mechanism could take the efficiency threshold $e_{\min}$ and the average channel conditions of the UE into account. Besides reducing the processing effort for MCS determination and MIMO mode selection, this would also reduce the overhead for CSI acquirement.

Summarizing, the newly introduced computational effort is considered to be low compared to the effort for encoding the data. In addition, the additional effort occurs in components which can cope with overload. Determining MCS and selecting MIMO mode are performed independently for each UE. This allows to parallelize the computations and to skip UEs in case of overload. The proposed prediction mechanism does not only take into account the effort for the encoding of the data, but also any other compute effort which delays the encoding. It thus manages the total effort, including that for the proposed system itself.

Besides the computational complexity, architectural complexity is also a relevant criterion. A benefit of the proposed system is that it does not introduce a component which requires synchronized coordination of a whole BBU pool. It does, however, assume that there is either a single compute unit, or a tightly synchronized pool of compute units which achieves ideal utilization of all resources. Suitable adaptation of the proposed system for a loosely coupled pool of compute units remains for further study.

#### *5.6.5.3 Interaction with Resource Allocation*

Section 5.6.1 stated that the proposed system should operate independently of RA. However, besides the alignment of the objectives of RA and the proposed system, which were discussed in section 5.6.3.6, the proposed system also interacts in two other ways with RA.

First, channel capacity predictions are distorted by changing the MIMO mode. Many RA algorithms rely on a prediction of the channel capacity experienced by the UEs. This prediction is often used as relative metric for opportunistic resource allocation. In that case, a false prediction of the capacity is tolerable if it impacts all UEs in a similar way. However, RA can also use the

[49] Evaluations in this thesis were all performed with $|\mathcal{M}| = 11$. See also section 5.4.3.2.

prediction to determine the number of PRBs required to deliver an absolute amount of data. This is necessary, e. g., for control messages of a fixed size, or in case only a limited amount of user data is to be transmitted. In these cases, it is inappropriate to reduce the capacity after RA has allocated PRBs, because few remaining bits force the system to allocate additional time-frequency resources in a consecutive subframe.

There are multiple approaches to avoid this interaction. RA could be enhanced to mark those UEs where the allocated capacity is strictly required and cannot be efficiently compensated by later resource allocations. The proposed MIMO mode selection heuristic is then applied only to the non-marked UEs. Another approach to handle this interaction is to modify the prediction of the channel capacity. That could already incorporate the expected MIMO mode selection, so that it better matches the realizable capacity. The RA can then assign more PRBs to those UEs to which it plans to deliver a fixed amount of data. The first approach reduces the efficiency of the proposed system, because the marked UEs use computationally less efficient MIMO modes. The second approach could cause undesired interactions and oscillations between RA and the proposed system, because the current value of $e_{\text{min}}$ then influences the UEs selected by RA. For the performance evaluations in this thesis, none of these approaches is implemented.

The second interaction discussed here is the canceling of allocated resources. As long as the fraction of canceled resources is low, this effect is dominated by decode errors caused by varying channel quality, and can therefore be ignored. However, in contrast to decode errors, canceled resources are immediately known at the eNodeB, and could therefore be compensated in the following subframe. Such compensation is not implemented for the evaluations performed here. For some messages, where immediate delivery is essential, robust encoding can be used to reduce the probability of decode errors. The RA mechanism can provide this information to the module which performs the encoding. The respective data can then be encoded first, so that it is not impacted by the skipping mechanism.

Summarizing, interaction with RA cannot be avoided completely. The described measures can be implemented to reduce possible negative impacts.

#### *5.6.5.4 Robustness*

As last requirement, robustness of the proposed system was requested in section 5.6.1. The main mechanism responsible for robustness of the proposed system is the fallback mechanism defined in section 5.6.3.3. This allows the evaluated building blocks of the system to cope with arbitrary low and / or sudden resource limits. Although the system can cope with such situations, efficient operation cannot be expected. This property is beneficial to tolerate hardware failures. In addition, it allows for operation in dynamic cloud environments, where an exact prediction of the resources is often difficult.

In addition to external influences, the fallback mechanism also serves to make the system robust w. r. t. false predictions of the processing effort. In case the model to estimate the compute effort for the MIMO modes is not exact, that results in undesired fluctuations of the resulting processing time. The fallback mechanism allows the system to handle this by skipping the encoding for some UEs. Thus, the inexact model impairs the efficiency of the system, but does not endanger its operation. A detailed quantitative evaluation of this property is out of scope of this thesis.

# 6 Performance Evaluation

The previous chapter presented a system which uses a combination of prediction and heuristics to cope with compute resource limitations in a BBU. Several aspects of the proposed system were already discussed in section 5.6.5. This chapter complements that with a thorough evaluation of the achieved performance.

The evaluation is split into two parts, which are based on two different system models. The first part compares the performance of the proposed system to that of an optimal solution. Thereby, the consequences of the simplifications performed during the design of the proposed system can be evaluated. However, this implicates that a simple system model is used, so that an all-encompassing optimization problem can be formulated and solved.

The second part amends the first one by an evaluation in a more realistic system model. It focuses on the effects of dynamic network load on the performance of the proposed system. These dynamic effects are expected to impact especially the prediction mechanism, which relies on the compute load of previous subframes to predict the efficiency threshold.

## 6.1 Evaluation Compared to Optimal Allocation

The objective of this section is to evaluate the performance achieved by the proposed system in a simple scenario. It is expected that it achieves lower network performance than the optimization problems used as basis for its design, because it is based on simplifications and heuristics. The extent of the impact of these simplifications is to be studied here.

In this section, the proposed system and a simple baseline heuristic are evaluated by simulation. As further references, multiple variants of an optimization problem are solved. These variants do successively limit the flexibility of the optimizer, and thereby allow to identify which simplification contributes to the lower performance achieved by the proposed system.

This section is structured as follows. First, section 6.1.1 describes how the evaluation methodology applied here differs from that used in the previous evaluations. Section 6.1.2 defines the optimization models used as reference and the baseline heuristic. The simulation and optimization studies are subsequently presented in section 6.1.3. Finally, the evaluations are concluded in section 6.1.4.

### 6.1.1 Evaluation Methodology

The main objective for the design of the evaluation methodology for these studies is that it shall be possible to compare the proposed system, evaluated by simulations, to optimal solutions for RA and MIMO mode selection, which are derived from optimization problems. To be suitable for optimization, the smaller scenario with seven sites and 21 sector cells, as introduced in section 5.3.2, is used. For the same reason, the full buffer traffic model is applied here.

The evaluation method applied here is similar to that described in section 5.5.1. Thus, only the differences are discussed here. To limit the scope of the evaluations, the system has been configured with 210 UEs, only. For the same reason, only the PF fairness scheme is studied.

The proposed system assumes the successive processing of subframes. Therefore, the approach of modeling a snapshot in time, as used for the optimization models in section 5.5, is not applicable here. Instead, for the evaluation in this section a limited period of simulated time of 100 ms is considered. This allows to formulate an all-encompassing optimization problem and also to perform a step by step simulation of the temporal behavior. To achieve a more realistic RA, the available system bandwidth of 10 MHz is split into 50 PRBs, which are assigned exclusively to the individual UEs. The correlation of the radio channels w. r. t. time and frequency is modeled according to the SCM as described in section 5.3.4.

To be able to solve the optimization problem with reasonable compute effort and memory, IfCo has been disabled for this evaluation. Thus, each BS is allowed to transmit on each PRB without considering the interference caused in neighbor cells. From the applied full-buffer traffic model results that each BS uses all PRBs for transmission. In case the proposed system cancels the resources allocated to a UE, the simulation model assumes that these PRBs still cause full interference to neighboring cells, i. e. it is assumed that useless data symbols are transmitted. This models the fact that neighboring BSs cannot adapt the used MCSs to make use of unused PRBs. Together with the assumption that the precoding actually used by interfering BSs does not influence the received interference (see section 5.3.5), this results in static interference.

For each parametrization, 20 independent UE drops are evaluated with a duration of 100 ms each. The simulation requires a warm-up phase, so that transient effects resulting from start conditions of the simulation do not influence the simulation results. Therefore, 10 s of simulated time have been prepended to each drop.[1] During this warm-up phase, the simulation operates but statistics are not updated. To obtain a stationary system state, a radio channel trace with the duration of 100 ms is read repeatedly. In contrast to the simulation, the studied optimization problems encompass the whole evaluated time of 100 ms and do not require a warm-up phase.

For RA in the simulations, using the PF heuristic as introduced in equations (3.11) and (3.12) in section 3.4.2 is a natural choice. This heuristic is not applicable for the optimization, because there RA is part of the optimization problem. We are here interested in the differences between the results of simulation and optimization. With different RA, it would be difficult to reason which effects are caused by the proposed system and which by different RA implementations. Therefore, the RA used for the simulations in this section is derived from the solution of an optimization problem. This is defined in the section 6.1.2.1.

[1] As large-scale channel effects and data traffic are constant, 10 s has shown to be sufficient to achieve a stationary state of the system.

As before, the evaluated metrics are the average UE rate and the 5th percentile of the UE rates. The network performance achieved by the proposed system is evaluated for different compute resource limits. It is compared to solutions to optimization problems, which serve as upper bound of the performance. In addition, it is contrasted with the performance achieved by a simple baseline heuristic. The reference configurations are presented in detail in section 6.1.2.

In addition to the absolute rates achieved with the different configurations, the relative differences of the rates compared to a reference configuration are evaluated. This allows to zoom into the range where the performance differs between the configurations. This evaluation is based on the fact that all configurations operate on the same problems, i. e. for each drop number, the positions and radio channels of the UEs are the same in all configurations. To calculate the relative differences, first, for each UE the achieved rate is divided by that achieved in the reference configuration, and 1 is subtracted from the quotient. Subsequently, the average of the resulting values is calculated over all UEs in all drops. This provides the evaluated metric.

The confidence intervals are calculated analogously to the method used for other metrics. Thereto, the average over the relative differences of all UEs in a drop is calculated. Following the central limit theorem, this drop-averages are assumed to be normally distributed. The t-distribution is then used to derive the confidence intervals from the drop-averages. The resulting intervals are smaller than those calculated for the average absolute rate, because variations caused by the radio channel are canceled out by comparing the same UE in different configurations.

The objective of these evaluations is to compare optimization and simulation. Thus, special care has been taken to achieve comparability in most components. For both realizations, the calculation of the small-scale effects of the radio channel and part of the PHY layer model are performed in *MATLAB*. A single *Java* program, which is based on the *IKR SimLib* [IKRSimLib] and *IKR RadioLib*, loads this data. It also loads the RA from a file which is generated by a preliminary optimization run. Repeatable streams of random numbers are used to ensure that the placement of UEs and other models give the same results for simulation and optimization. The *Java* program then either runs the simulator or defines and configures the optimization problem. The latter is solved by the *IBM ILOG CPLEX Optimizer*.

There is a single aspect which is not aligned between simulation and optimization: The simulation model allows each UE to be served with one or two codewords per subframe (depending on the number of spatial layers), where each of these codewords is transmitted with a single MCS. This constraint is not kept for the optimization. To be able to solve the optimization problems with feasible resources, the selection of the MCSs cannot be included in the problem formulations. Instead, separate MCSs are precalculated for each PRB. The optimization problems therefore allow that a single UE is served on multiple PRBs with different MCSs in one subframe.

The remaining aspects of the evaluation methodology are the same as introduced in section 5.5.1. This comprises the parametrization of the optimization solver.

### 6.1.2 Reference Configurations

For the studies in this section, four optimization problems are evaluated as reference. These problems encompass the whole evaluated time and frequency range. The first problem, which is defined in section 6.1.2.1, allows the optimizer to flexibly allocate time-frequency resources and select MIMO modes.[2] A variant of this problem is also used to decide the allocation of time-frequency resources to UEs for the remaining problems and for the simulation studies.

Compared to the first problem, the proposed system has several limitations by design. The second problem adapts only MIMO modes to the available compute resources. It serves to evaluate the impact of this limitation on the network performance. This problem is defined in section 6.1.2.2. Two other optimization problems are defined in section 6.1.2.3. These model the fact that the proposed system does not allocate MIMO modes optimally. Instead, multiple simplifications were made to design a simple distributed heuristic.

Finally, a simple baseline heuristic is introduced in section 6.1.2.4. This shows the performance achievable without MIMO mode adaptation.

#### *6.1.2.1 Optimal Resource Allocation and Link Adaptation*

The applied resource model is similar to that defined in section 5.4.3.1. That defines power allocation resources to model the fact that neighboring BSs transmit with either zero or full power, which influences the received interference. These are here replaced by channel resources, which resemble different PRBs with individual channel characteristics. To formulate a more realistic RA, PRBs are considered to be indivisible, and each PRB can be assigned to a single UE, only. Consequently, the sizes of the channel resources are fixed. In addition, each scheduling resource has either the size zero or has the same size as the corresponding channel resource, i. e. one PRB.

As before, different MIMO modes can be used to serve the UEs. However, to be more realistic and to apply the same constraints in the simulation and the optimization, in each subframe each UE can only be served with a single MIMO mode. Therefore, sizes of the MIMO resources are also either zero or one PRB, and an additional constraint prohibits the combination of different modes for one UE and subframe. These differences suggest to use a formulation based on binary flag variables instead of continuous sizes.

Assume that the set of all channel resources $\mathcal{R}^{\mathrm{c}}$ resembles a two-dimensional structure of PRBs. It is defined as

$$\mathcal{R}^{\mathrm{c}} = \left\{ n^{\mathrm{c}}_{s,f} \colon s \in \mathcal{S}, f \in \mathcal{F} \right\}, \tag{6.1}$$

where $\mathcal{S}$ is the set of subframes and $\mathcal{F}$ (with $|\mathcal{F}| = N_{\mathrm{PRB}}$) the set of all PRB positions in frequency dimension. The channel resources of a single subframe $s$ can be grouped as

$$\mathcal{R}^{\mathrm{c}}_{s} = \left\{ n^{\mathrm{c}}_{\bar{s},f} \in \mathcal{R}^{\mathrm{c}} \colon \bar{s} = s \right\} \quad \forall s \in \mathcal{S}. \tag{6.2}$$

The allocation of scheduling resources to UEs and the selection of MIMO modes is represented by binary variables $a_{u,n^{\mathrm{c}},m}$ with $u \in \mathcal{U}$, $n^{\mathrm{c}} \in \mathcal{R}^{\mathrm{c}}$, and $m \in \mathcal{M}$. The value of $a_{u,n^{\mathrm{c}},m}$ is one if the channel resource $n^{\mathrm{c}}$ is used to serve UE $u$ with MIMO mode $m$ and zero otherwise. An additional set of binary variables $\tilde{a}_{u,s,m}$ (with $u \in \mathcal{U}$, $s \in \mathcal{S}$, and $m \in \mathcal{M}$) is defined to prohibit

[2] In contrast to the problems defined in section 5.4, IfCo is not considered here.

the combination of different MIMO modes. A variable $\tilde{a}_{u,s,m}$ is one if UE $u$ uses MIMO mode $m$ in subframe $s$ and zero otherwise.

For each UE, channel resource, and MIMO mode the MCS delivering maximum throughput can be selected.[3] Based on this, the resulting data capacities and processing efforts can be calculated in advance. These are here denoted as $r_{u,n^{\mathrm{c}},m}$ and $p_{u,n^{\mathrm{c}},m}$, respectively. The total amount of data received by an individual UE $u$ is then

$$r_u = \sum_{n^{\mathrm{c}}\in\mathcal{R}^{\mathrm{c}}} \sum_{m\in\mathcal{M}} a_{u,n^{\mathrm{c}},m} r_{u,n^{\mathrm{c}},m}. \tag{6.3}$$

Based on these definitions, the basic problem is defined as follows.

$$\max_{\substack{a_{u,n^{\mathrm{c}},m},\\ u\in\mathcal{U},n^{\mathrm{c}}\in\mathcal{R}^{\mathrm{c}},m\in\mathcal{M},\\ \tilde{a}_{u,s,m},\\ u\in\mathcal{U},s\in\mathcal{S},m\in\mathcal{M}}} \quad \sum_{u\in\mathcal{U}} \mathrm{U}_{\mathrm{cap}}(r_u) \tag{6.4a}$$

$$\text{s. t.} \quad \sum_{u\in\mathcal{U}} \sum_{n^{\mathrm{c}}\in\mathcal{R}^{\mathrm{c}}_s} \sum_{m\in\mathcal{M}} a_{u,n^{\mathrm{c}},m} p_{u,n^{\mathrm{c}},m} \le p^{\mathrm{abs}}_{\max} \qquad \forall s\in\mathcal{S} \tag{6.4b}$$

$$\sum_{u\in\mathcal{U}_b} \sum_{m\in\mathcal{M}} a_{u,n^{\mathrm{c}},m} = 1 \qquad \forall b\in\mathcal{B}, n^{\mathrm{c}}\in\mathcal{R}^{\mathrm{c}} \tag{6.4c}$$

$$\sum_{m\in\mathcal{M}} \tilde{a}_{u,s,m} = 1 \qquad \forall u\in\mathcal{U}, s\in\mathcal{S} \tag{6.4d}$$

$$a_{u,n^{\mathrm{c}},m} \le \tilde{a}_{u,s,m} \qquad \forall u\in\mathcal{U}, s\in\mathcal{S}, n^{\mathrm{c}}\in\mathcal{R}^{\mathrm{c}}_s \tag{6.4e}$$

$$a_{u,n^{\mathrm{c}},m} \in \{0,1\} \qquad \forall u\in\mathcal{U}, n^{\mathrm{c}}\in\mathcal{R}^{\mathrm{c}}, m\in\mathcal{M} \tag{6.4f}$$

$$\tilde{a}_{u,s,m} \in \{0,1\} \qquad \forall u\in\mathcal{U}, s\in\mathcal{S}, m\in\mathcal{M} \tag{6.4g}$$

In constraint (6.4b), $p^{\mathrm{abs}}_{\max}$ denotes the absolute compute resource limit per subframe. It is defined as $p^{\mathrm{abs}}_{\max} = P^{\mathrm{peak}}_{|\mathcal{B}|N_{\mathrm{PRB}}} p_{\max}$.

This problem simultaneously adapts resource allocation and MIMO mode selection, so that the sum utility is maximized under limited time-frequency and compute resources. Here, again, the PF utility function as defined in equation (5.52) is applied. Constraint (6.4b) limits the total processing effort for each subframe. Equation (6.4c) ensures orthogonality of resource allocation within each BS. Constraints (6.4d) and (6.4e) prohibit the usage of multiple MIMO modes to serve a single UE at one subframe. Finally, the binarity of the variables is enforced by equations (6.4f) and (6.4g). In the evaluations, this problem is denoted as *opt. RA and LA*.

A derived form of this problem is also used for resource allocation in the simulation runs. Thereto, unlimited compute resources are assumed, i. e. $p^{\mathrm{abs}}_{\max} = \infty$. This effectively removes constraint (6.4b). The solution to this modified problem resembles the optimal resource

[3] Note that this is different to the simulation, where a single MCS is used for all PRBs allocated to a UE in a subframe. However, capturing this constraint would have made the formulation of the optimization problem even more complex.

allocation without considering limited compute resources. For the simulation studies, only the allocation of resources to UEs is extracted from the optimal solution, while the MIMO modes are ignored. PRB allocation flags can be derived from the solution. These are here defined as $a^{(\infty)}_{u,n^{\mathrm{c}}} = \sum_{m\in\mathcal{M}} \widehat{a_{u,n^{\mathrm{c}},m}}\ \forall u \in \mathcal{U}, n^{\mathrm{c}} \in \mathcal{R}^{\mathrm{c}}$, where $a^{(\infty)}_{u,n^{\mathrm{c}}} \in \{0,1\}$ and $\widehat{a_{u,n^{\mathrm{c}},m}}$ denotes the value of the respective binary variable in the optimal solution to problem (6.4). These PRB allocation flags are taken as input for the simulation.

#### 6.1.2.2 Optimal Link Adaptation

The problem (6.4) allows the optimizer to simultaneously adapt RA and MIMO mode selection. In contrast, the proposed system only adapts MIMO modes. To assess the impact of this limitation on the network performance, the following problem *opt. LA* is defined.[4]

Assume that problem (6.4) has been solved for $p^{\mathrm{abs}}_{\mathrm{max}} = \infty$. Let $a^{(\infty)}_{u,n^{\mathrm{c}}}\ \forall u \in \mathcal{U}, n^{\mathrm{c}} \in \mathcal{R}^{\mathrm{c}}$ denote the allocation of channel resources to UEs in the optimal solution. Then, modify problem (6.4) by enforcing the same resource allocation as in that solution. Thereto, a constraint $a^{(\infty)}_{u,n^{\mathrm{c}}} = \sum_{m\in\mathcal{M}} a_{u,n^{\mathrm{c}},m}\ \forall u \in \mathcal{U}, n^{\mathrm{c}} \in \mathcal{R}^{\mathrm{c}}$ can be added. From the definition and interrelationship of the variables in problem (6.4) follows that this is equivalent to $a_{u,n^{\mathrm{c}},m} = a^{(\infty)}_{u,n^{\mathrm{c}}}\, \tilde{a}_{u,s,m}\ \forall u \in \mathcal{U}, s \in \mathcal{S}, n^{\mathrm{c}} \in \mathcal{R}^{\mathrm{c}}_s$. This allows to omit the resource allocation variables from the problem. Equation (6.3) can thus be reformulated as

$$r_u = \sum_{s\in\mathcal{S}} \sum_{m\in\mathcal{M}} \tilde{a}_{u,s,m} \sum_{n^{\mathrm{c}}\in\mathcal{R}^{\mathrm{c}}_s} a^{(\infty)}_{u,n^{\mathrm{c}}}\, r_{u,n^{\mathrm{c}},m}. \tag{6.5}$$

With these definitions, the problem *opt. LA* is just concerned with selecting MIMO modes per UE and subframe. It is defined as

$$\max_{\substack{\tilde{a}_{u,s,m},\\ u\in\mathcal{U}, s\in\mathcal{S}, m\in\mathcal{M}}} \quad \sum_{u\in\mathcal{U}} \mathrm{U}_{\mathrm{cap}}(r_u) \tag{6.6a}$$

$$\text{s. t.} \quad \sum_{u\in\mathcal{U}} \sum_{m\in\mathcal{M}} \tilde{a}_{u,s,m} \sum_{n^{\mathrm{c}}\in\mathcal{R}^{\mathrm{c}}_s} a^{(\infty)}_{u,n^{\mathrm{c}}}\, p_{u,n^{\mathrm{c}},m} \le p^{\mathrm{abs}}_{\mathrm{max}} \quad \forall s \in \mathcal{S} \tag{6.6b}$$

constraints (6.4d) and (6.4g).

Here, the compute capacity constraint (6.6b) is equivalent to the constraint (6.4b), but reformulated to remove the resource allocation variables.

#### 6.1.2.3 Optimal Efficiency Threshold

The previously defined problem *opt. LA* just adapts MIMO modes to cope with limited processing capacity. However, to do so it can freely combine different MIMO modes for different UEs. In contrast, the proposed system is bound to define a single efficiency threshold value, which then

[4] Note that this is similar to the evaluations in section 5.5.4, however a different scenario is applied here and the results will be compared to the performance achieved by the proposed system.

determines which UE uses which MIMO mode. This is equivalent to using the greedy algorithm to solve the knapsack problem as described in section 5.6.2.2. In addition, the proposed prediction mechanism assumes that the same value of the threshold is valid for consecutive subframes. A third optimization problem is defined in this section, which serves to evaluate the impact of these two limitations. Two variants model the case where the threshold is adapted individually per subframe and the case where it is constant. These variants are denoted as *opt.* $e_{\min}$ *(per subframe)* and *opt.* $e_{\min}$ *(constant)*, respectively. The threshold values are then used to derive the MIMO modes for all subframes. Both use the allocation of channel resources as optimized for unlimited compute resources.

The proposed system skips the encoding of data for some UEs whenever too complex MIMO modes are selected. It is acceptable that the optimal value of the threshold causes overload in some subframes, because that can be required to achieve high utilization of compute resources. However, if the solver were allowed to decide which UEs do not get served, it would have an additional degree of freedom that is not available to the proposed system. Therefore, the skipping of UEs is here not modeled directly in the optimization problem. Instead, the optimizer is allowed to proportionally scale compute effort and data capacity of all UEs served in a subframe. This models the fact that the skipping hits all UEs with the same probability.

The variant *opt.* $e_{\min}$ *(per subframe)* is defined as follows. Assume that problem (6.4) has been solved for $p_{\max}^{\text{abs}} = \infty$. As before, let $a_{u,n^{\text{c}}}^{(\infty)}$ $\forall u \in \mathcal{U}, n^{\text{c}} \in \mathcal{R}^{\text{c}}$ denote the allocation of channel resources to UEs in the optimal solution. Based on that resource allocation, let dominated and LP-dominated modes be determined as described in section 5.6.2.2. Let $\widetilde{\mathcal{M}}_{u,s}$ denote the modes of UE $u$ which are neither dominated nor LP-dominated in subframe $s$. Further, assume that the modes of each set are sorted by utility and the corresponding indices assigned. Define the efficiency $e_{u,s,m}$ of all except the least complex mode in each set to be calculated as in equation (5.47).

Let $e_{\min,s}^{\text{opt}}$ be the variable denoting the optimal efficiency threshold in subframe $s$ ($s \in \mathcal{S}$). This is used to select the MIMO modes, which are variables in the previous problems. For each useful mode $m$ define an interval $(e_{u,s,m}^{(\text{-})}, e_{u,s,m}^{(+)}]$ of efficiency threshold values where it is active. The least complex mode is active whenever the efficiency threshold is arbitrarily large. However, it is not used any more when the threshold is sufficiently small to enable the next complex mode. Thus, for the least complex mode $e_{u,s,m_0}^{(\text{-})} = e_{u,s,m_1}$ and $e_{u,s,m_0}^{(+)} = \infty$. Similarly, the other modes except the most complex have their interval bounded by their own efficiency and that of the respective next complex mode in the set. This results in $e_{u,s,m_i}^{(\text{-})} = e_{u,s,m_{i+1}}$ and $e_{u,s,m_i}^{(+)} = e_{u,s,m_i}$. Finally, the most complex mode is active whenever the threshold is lower than its efficiency. Consequently, $e_{u,s,m}^{(\text{-})} = -\infty$ and $e_{u,s,m}^{(+)} = e_{u,s,m}$.

Furthermore, let $\breve{a}_s$ with $s \in \mathcal{S}$ denote the set of scaling variables. This scales the UE rates in each subframe, so equation (6.5) is modified to

$$r_u = \sum_{s \in \mathcal{S}} \breve{a}_s \sum_{m \in \mathcal{M}} \tilde{a}_{u,s,m} \sum_{n^{\text{c}} \in \mathcal{R}_s^{\text{c}}} a_{u,n^{\text{c}}}^{(\infty)} r_{u,n^{\text{c}},m}. \tag{6.7}$$

This allows to define *opt.* $e_{\min}$ *(per subframe)* as follows.

$$\max_{\substack{e^{\text{opt}}_{\text{min},s}, s\in\mathcal{S},\\ \breve{a}_s, s\in\mathcal{S}}} \sum_{u\in\mathcal{U}} \mathrm{U}_{\text{cap}}(r_u) \tag{6.8a}$$

$$\text{s. t.} \quad \tilde{a}_{u,s,m} = \begin{cases} 1 & \text{if } e^{\text{opt}}_{\text{min},s} > e^{(\text{-})}_{u,s,m} \wedge e^{\text{opt}}_{\text{min},s} \leq e^{(+)}_{u,s,m} \\ 0 & \text{otherwise} \end{cases} \quad \forall u\in\mathcal{U}, s\in\mathcal{S}, m\in\widetilde{\mathcal{M}}_{u,s} \tag{6.8b}$$

$$\tilde{a}_{u,s,m} = 0 \quad \forall u\in\mathcal{U}, s\in\mathcal{S}, m\in\mathcal{M}\setminus\widetilde{\mathcal{M}}_{u,s} \tag{6.8c}$$

$$\breve{a}_s \sum_{u\in\mathcal{U}} \sum_{m\in\mathcal{M}} \tilde{a}_{u,s,m} \sum_{n^{\text{c}}\in\mathcal{R}^{\text{c}}_s} a^{(\infty)}_{u,n^{\text{c}}} p_{u,n^{\text{c}},m} \leq p^{\text{abs}}_{\text{max}} \quad \forall s\in\mathcal{S} \tag{6.8d}$$

$$0 \leq \breve{a}_s \leq 1 \quad \forall s\in\mathcal{S} \tag{6.8e}$$

Here, the objective (6.8a) is based on the rate calculation as defined in equation (6.7). Constraint (6.8b) enforces the MIMO modes to be selected following the definition of the efficiency intervals. Equation (6.8c) ensures that dominated and LP-dominated modes are not used. Constraint (6.8d) limits the processing effort per subframe and thereby considers the scaling variables. Finally, equation (6.8e) confines the range of the scaling variables.

The second variant *opt.* $e_{\text{min}}$ *(constant)* constrains the threshold to be the same in all subframes. Thereto, the variable $e^{\text{opt}}_{\text{min}}$ is introduced. The problem is defined as follows.

$$\max_{\substack{e^{\text{opt}}_{\text{min}}\\ \breve{a}_s, s\in\mathcal{S}}} \sum_{u\in\mathcal{U}} \mathrm{U}_{\text{cap}}(r_u) \tag{6.9a}$$

$$\text{s. t.} \quad e^{\text{opt}}_{\text{min},s} = e^{\text{opt}}_{\text{min}} \quad \forall s \in \mathcal{S} \tag{6.9b}$$

and equations (6.8b), (6.8c), (6.8d), and (6.8e)

#### *6.1.2.4 Baseline Heuristic*

In addition to the optimization problems used as reference, the following studies also evaluate the performance of a simple baseline heuristic, denoted as *baseline*. The objective of this heuristic is to show the performance achievable without adapting MIMO modes. To still be able to cope with limited processing resources, it operates as follows.

IfCo, RA, and LA are performed without considering the available resources. Thus, MIMO modes are selected to maximize the spectral efficiency for the respective set of allocated PRBs. Subsequently, system starts to encode the data for all UEs with allocated PRBs in random order. When overload occurs, i. e. the time provided for encoding is over before data for all UEs could be encoded, the remaining UEs are skipped. Skipped UEs receive no data in the current subframe.

This heuristic is equivalent to the fallback mechanism of the proposed system, which was defined in section 5.6.3.3. Compared to the schematic picture of the proposed system in figure 5.19, it contains only the modification to step D, but neither step H nor the modification to step C.

Note that the decision to skip the encoding for a UE comes only when the data for the remaining UEs has already been encoded. Therefore, in reality this cannot be taken into account for the

MCS selection in neighboring cells. However, if the fraction of skipped encodings becomes significant, the lower average interference could allow to decode less robustly encoded data. The efficiency then depends on the measurement of CQI and on the accepted error probability. This cannot be directly integrated into the simulation model, because that assumes ideal CQI.

There are two approaches to simplify this effect on the interference. First, it could be assumed that the interference level is not reduced by skipped resources. This is equivalent to a system that transmits random data on the respective symbols. Second, it could be assumed that the information about skipped encodings can be taken into account for the MCS selection in neighboring cells. That is not possible in reality, because the data to be transmitted is already encoded when the skipped encodings are known. It can, however, be implemented in the simulation model by iteratively selecting transmissions to be skipped and re-encoding the data.[5]

The first approach is the same which is also used to evaluate the proposed system. However, it under-estimates the achievable network performance. Therefore, the second approach is used for the evaluation of the baseline heuristic. Consequently, the simulated performance of the baseline heuristic is not realizable in reality.

### 6.1.3 Studies

The previous sections defined the evaluation methodology and the reference configurations to be applied for the studies. This section presents the evaluation results. Its central objective is to evaluate the impact of the simplifications made in the design of the proposed system on the network performance. The proposed system is configured with $p_{\text{off}} = 0.25\,\%$ and $e_{\text{step}} = 0.1$. The tuning studies presented in appendix C have shown this to be efficient configurations.

#### *6.1.3.1 General Performance*

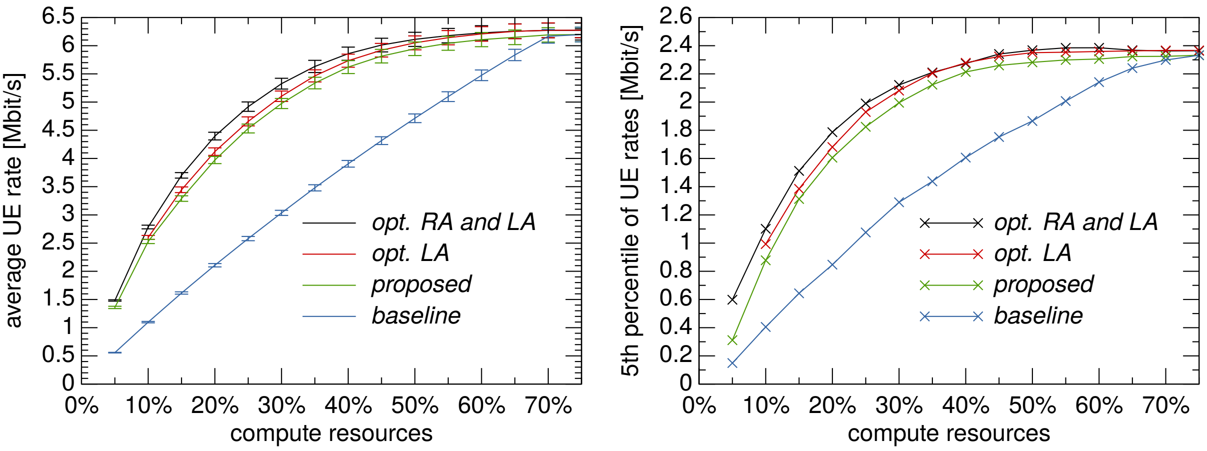
Figures 6.1a and 6.1b show plots with the average UE rate and the 5th percentile of the UE rate on the $y$-axis, respectively. As before, the available compute resources are plotted on the $x$-axis. The curves represent the reference optimization problems *opt. RA and LA* and *opt. LA*, the proposed system, and the baseline heuristic. They follow the general shape of those presented in section 5.5.

For compute resources above 70 %, network performance is not impacted. In this range, no significant difference between the compared problems and heuristics is expected. Indeed, the optimized solutions show the same network performance. Compared to that, the simulations achieve 1.2 % reduced average UE rate (1.8 % for 5th percentile of the UE rates). This can be explained as the penalty for using a single MCS per UE and subframe.[6]

For lower compute resource limits, a larger difference of the compared systems is expected, because there the different approaches to cope with limited processing capacity come into play.

[5] Adapting the MCSs to reduced interference can result in more complex encoding. Thus, it can be necessary to skip more resource allocations. This process is repeated until the compute capacity is not exceeded any more.

[6] As stated in section 6.1.1, this is the only difference between the optimization problems and the simulations for the configuration with unlimited compute resources.



**(a)** average UE rate **(b)** 5th percentile of UE rates

**Figure 6.1:** Evaluation of the UE rate

Here, the average rates achieved by the proposed system and by the two optimization problems do not differ significantly. At the same time, the baseline heuristic suffers from an almost linear decrease of the average rate as soon as compute resources fall below 70 %.

The radio network requires between 70 % and 75 % of the theoretical peak compute resources to operate without impacting the network performance. With only half of this, i. e., about 35 %, the average rate of the proposed system is still 86 % of the rate achieved in the unlimited case. For the same compute resource limit, the baseline heuristic maintains only 56 % of the original average rate. Any other approach which is restricted to adapt MIMO modes (modeled by *opt. LA*) cannot achieve more than 87 %. When also incorporating changes to RA (modeled by *opt. RA and LA*), not more than 90 % are possible.

The 5th percentile plotted in figure 6.1b behaves similarly as the average rates. Especially, the proposed system does not significantly impact cell border UEs to achieve better average performance. Comparing the average and the percentile achieved by the *baseline* configuration, the benefit of reduced interference for cell border UEs becomes visible.

The proposed system skips encoding of data in case the predicted threshold results in overload. For 15 % to 65 % compute resources, about 0.5 % to 0.8 % of the allocated resources get skipped (not shown in the plots). This value is much larger in case even the simplest MIMO modes are too complex to meet the compute resource limit. Thus, in the simulations performed with $p_{\max} = 5\,\%$, about 42 % of the allocated resources are skipped.

#### 6.1.3.2 *Performance Drop Compared to Optimal Solution*

To better illustrate the causes for the performance drop of the proposed system, the average relative difference of the UE rates is plotted in figure 6.2. Thereto, for each UE, the relative drop of the rate compared to the rate achieved by *opt. RA and LA* is calculated. For each parametrization, the plot shows the average over the relative differences of all UEs.[7]

[7] The method used to calculate confidence intervals for these values is described in section 6.1.1.

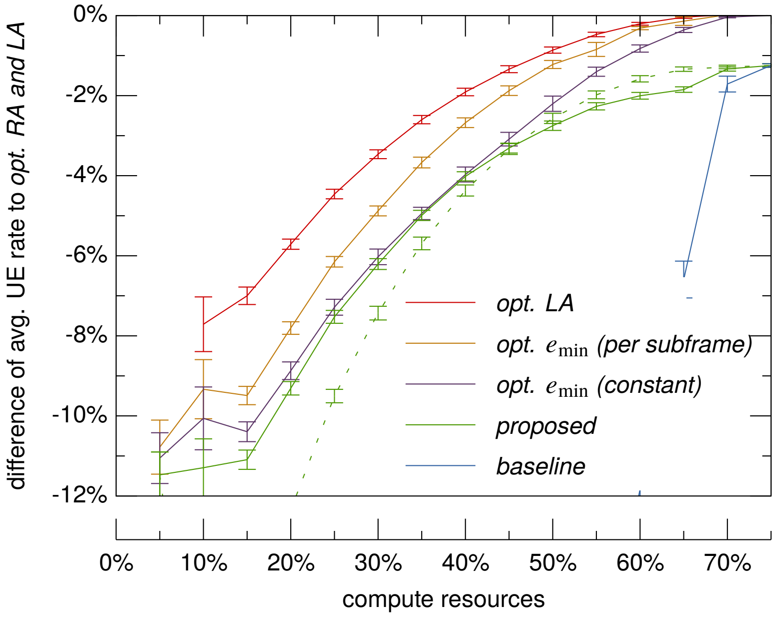


**Figure 6.2:** Drop of average UE rate

In addition to the configurations from figure 6.1, the two additional reference configurations *opt.* $e_{\min}$ *(per subframe)* and *opt.* $e_{\min}$ *(constant)* are shown here. Also, two different parametrizations of the proposed system are shown. The solid line, marked as *proposed*, represents the configuration $p_{\text{off}} = 0.25$ %. The dotted line represents $p_{\text{off}} = 1$ %.

Note that it is difficult to cope with compute resource limits of 5 % and 10 % by adapting MIMO modes alone. Thus, for these parametrizations the skipping and scaling mechanisms of the compared configurations have significant influence on the network performance. For example, the reference configuration *opt. RA and LA* here utilizes only 55 % and 94 % of the available PRBs. With 5 % compute resources, the optimization problem *opt. LA* becomes infeasible, so the respective point is missing from the chart. At the same parametrization, the proposed system skips encoding of 42 % of the allocated PRBs. The remainder of the discussion focuses on higher compute resource limits, which are assumed to be more common operating points of the system.

The following paragraphs follow the structure of the discussion in section 5.6.5.1. That lists four causes for performance drops. First, the proposed system considers each subframe separately, which is suboptimal compared to a joint optimization over all consecutive subframes. The following comparisons show that the other three causes fully explain the performance drop of the proposed system. Thus, separate evaluation of each subframe has no significant impact on the performance.

The second cause for performance drops is the restriction to adapting only MIMO modes. By also modifying RA, the problem *opt. RA and LA* balances compute resource utilization between subframes. It also avoids allocations of PRBs which can only be used efficiently with complex MIMO modes. Figure 6.2 shows that, compared to *opt. RA and LA*, the performance of *opt. LA* drops by up to 7.7 %. The impact of this simplification increases for more stringent compute resource limits.

The third reason for reduced performance is that the proposed system does not select MIMO modes optimally. Instead, it uses a greedy heuristic to solve the MIMO mode selection problem.

This is reflected by the reference configuration *opt.* $e_{\min}$ *(per subframe)*. Compared to *opt. LA*, this further reduces the performance for low compute resource limits. There, it entails an additional performance penalty of 2.5 %. This optimization problem also defines a scaling variable per subframe. However, that is not used by the optimizer, i. e., $\breve{a}_s = 0$, except for $p_{\max} < 10\,\%$. There, the compute resource limit cannot be met otherwise.

The fourth cause for performance impairments is the prediction of the threshold $e_{\min}$. Section 5.6.5.1 further subdivides this into three different effects. The most influential of these are here fluctuations in the radio channel and actions of the RA, which cause variations of the compute load even if the load of the network is constant.[8] The proposed system is incapable of anticipating these variations. This limitation is evaluated by the reference configuration *opt.* $e_{\min}$ *(constant)*. It models the assumption that a single threshold is applicable for the evaluated time. When comparing *opt.* $e_{\min}$ *(constant)* to *opt.* $e_{\min}$ *(per subframe)*, the largest difference occurs at medium compute resource limits. At 40 % compute resources it reaches 1.3 %.

In the optimal solutions to this problem, the average value of the scaling variables $\breve{a}_s$ is equal to one for compute resources of 50 % and higher (not shown in the plots). The value falls below 0.97 only for 5 % compute resources, where operation of the system cannot be realized with the simplest MIMO modes alone. This implies that a parametrization which avoids skipping is beneficial for most resource limits.

Compared to the optimization problems, the average rate achieved by the proposed system is reduced by 1.2 % for practically unlimited compute resources. As stated before, this can be explained by the different modeling of the MCSs in the simulator. This difference diminishes when the compute resource limit is lowered. For compute resources between 25 % and 40 %, the performance achieved by the proposed system is almost the same as that achieved by *opt.* $e_{\min}$ *(constant)*.

As also shown in appendix C, the offset parameter $p_{\text{off}}$ can be used to tune the performance for different compute resource limits. The parametrization $p_{\text{off}} = 1\,\%$ (the dotted line) achieves higher performance for $p_{\max} > 45\,\%$, but suffers a significant performance drop for low compute resource limits. This aligns with the optimal values of the scaling variables $\breve{a}_s$ in the problem *opt.* $e_{\min}$ *(constant)*. A more elaborate prediction mechanism can potentially achieve the performance of the respective better parametrization in both ranges. However, exceeding the performance of *opt.* $e_{\min}$ *(constant)* is not possible without anticipating the fluctuations of the radio channel and the actions of the RA algorithm.

When comparing the proposed system directly to *opt. RA and LA*, the relative drop reaches 11 % at 5 % compute resources. However, the performance of the proposed system is better for higher compute resource limits. As long as more than half of the compute resources required for non-impaired operation are available (i. e. more than 35 %), the performance of the proposed system drops not by more than 5 % compared to the optimum. Thus, it allows to efficiently handle compute resource overload as long as that is not too extreme. For lower compute resource limits, the bigger part of the performance drop is caused by the restriction to only adapt MIMO modes. To avoid this drop, RA and LA have to be adapted simultaneously.

[8] The remaining two effects are that the control loop requires frequent overload to react and variations caused by dynamic load of the network. The former can almost be avoided by configuring a sufficiently large value for the offset $p_{\text{off}}$. The latter is not modeled in this evaluation.

### 6.1.4 Summary and Conclusion

The evaluations presented in this section compared the proposed system to a simple baseline heuristic and to different variants of an optimization problem. To facilitate this comparison, a simplifying full-buffer traffic model was used.

The results discussed in section 6.1.3.1 indicated that in this scenario the performance of the proposed system is close to the global optimum. By adapting MIMO modes, it can maintain high network performance when the available compute resources get limited. With only 50 % compute resources, it still achieves more than 95 % of the average data rate. Under the same conditions, the baseline heuristic suffers from a linear degradation of the network performance. At some compute resource limits the baseline heuristic cannot maintain half of the average rate achieved by the proposed system.

Section 6.1.3.2 zoomed into the differences between the optimal solution and the outcome of the proposed heuristic. It compared the results of different variants of the optimization problem, which successively limit the flexibility of the optimizer to approach the proposed system. It has shown that the performance is close to the optimum in case the compute resources are marginally restricted. The difference increases as the compute resource limit becomes tighter. The bigger part of the performance drop is induced with the limitation to adapt only MIMO modes.

The proposed system performs well in the applied evaluation scenario. Especially in case the compute resources are only slightly overbooked, the network performance is barely impacted.

It was stated before that adaptation of the RA mechanism is avoided, because that would cause additional complexity. Given that restriction, another heuristic cannot perform significantly better than the proposed one. Lifting this restriction could allow to increase the average rate by 5 % to 10 % for situations with stringent compute resource limits. It is, however, assumed that this is not worth the effort of modifying the RA mechanism.

## 6.2 Evaluation with Dynamic Load

The evaluations in the previous section have shown that the proposed system allows an LTE BBU pool to cope with limited processing resources without significantly impacting network performance. However, those evaluations were performed with a simplifying full-buffer traffic model to allow comparison with an optimization problem. That model implies that the system is fully loaded and that there are only little variations in the requested compute capacity.

The proposed system assumes that the efficiency threshold can be predicted based on the compute effort which occurred in previous subframes. It therefore requires a correlation of compute effort of consecutive subframes. The higher the variations of the compute load is, the more difficult the prediction of the required efficiency becomes. The main objective of this evaluation is to show whether a simple prediction mechanism such as the proposed one is applicable for a system with realistic dynamic behavior.

Furthermore, the full-buffer traffic model used for the evaluations in section 6.1 causes a constantly high load of the radio network. In reality, the load is lower and more variable. While this makes prediction of the compute load more difficult, it also reduces the average compute requirements. In addition, the fluctuating load suggests that it is possible to balance the compute load over consecutive subframes. Evaluating the effect of this on the network performance achieved with the proposed system is a second objective of the following studies.

This section is structured as follows. First, section 6.2.1 discusses the effects which can cause variations of the compute load. Subsequently, section 6.2.2 describes the applied evaluation methodology, which includes the system model and the reference configurations. Initial calibration of the model is performed in section 6.2.3 based on pilot runs. Section 6.2.4 presents the evaluation results, and section 6.2.5 concludes with a discussion of the results.

### 6.2.1 Effects Causing Variations of the Compute Load

It is important for the model used in this section to capture all relevant effects which cause variations in the compute effort. The external influencing factors for the actions of a mobile communication system are the data traffic, the channel, and the system configuration. Changing system configuration, e. g. by operating personnel or by SON mechanisms, is not considered here. The expected effects from data traffic and channel are discussed in the following paragraphs.

The data traffic also has a significant influence on the accruing compute effort. For example, users starting a new transmission can have significantly different channel conditions than other users. In addition, a new request can cause the system to transmit on previously free resources. A majority of the data transmissions can be assumed to be small.[9] Thus, in a significantly loaded system there is a large number of small transmissions. This results in many newly starting transmissions and many transmissions being completed per time interval. The effects from dynamic data traffic are considered as relevant and are therefore included in the evaluations in this section.

Channel variations are caused by mobility of the users and changes in the environment (e. g. movement of shadowing or reflecting surfaces). Following the introduction in section 5.3.4, these effects can be separated into macro-scale attenuation and small-scale fading effects. The small-scale fading effects were already regarded in the evaluations in section 6.1 and are also included in the evaluations in this section. In contrast, slow fading effects do not cause significant changes between consecutive subframes.[10] They are therefore not modeled here.

The channel also determines the serving cell of a UE. When a UE moves and the channel changes as result thereof, it can be handed over to a different cell. This instantaneously changes the load in the two involved cells. However, handover events cause the same kind of load variations as begin and end of data transmissions, and those data traffic events occur more frequently. Effects from handover are therefore omitted here.

In urban scenarios where cells are placed densely, interference also significantly influences the channel capacity. As a consequence of data traffic changes and RA decisions in neighboring cells,

[9] For example, in the data traffic model used here, 29 % of the data traffic objects are smaller than 10 KiB, and 58 % are smaller than 100 KiB. See also figure 5.4.

[10] According to [3GPP 25.913], LTE is designed to support velocities up to 500 km/h, while high performance should be maintained for up to 120 km/h. This corresponds to a movement of 14 cm and 3 cm per subframe, respectively. For comparison, the correlation distance for shadow fading is typically modeled to be 50 m (see also section 5.3.4).

the set of applicable MIMO modes can change quickly. This determines the accruing compute effort. Hence, the effect from interference is included in these evaluations.

Summarizing, the most relevant effects which cause variations in the compute load are assumed to be the data traffic and the interference. In addition, the RA algorithm influences compute load and interference. Consequently, these are modeled for the evaluations in this section.

### 6.2.2 Evaluation Methodology

The evaluation is performed by simulation. This is similar to the previous evaluations, but uses a more complex system model. Its parametrization is described in section 6.2.2.1. The different model also implicates adaptation of the evaluated metrics, which are introduced by section 6.2.2.2. An all-encompassing optimization problem cannot be used as reference. Instead, a smaller optimization problem is embedded into the simulation. This is defined in section 6.2.2.3. As the applied RA algorithm influences the compute load and significantly contributes to the performance of the system, it has to be modeled realistically. Section 6.2.2.4 describes the variant of the PF heuristic, which is used here. The implementation of the simulation model and the procedure of the simulation are illustrated in sections 6.2.2.5 and 6.2.2.6, respectively.

#### *6.2.2.1 Parametrization of the System Model*

The system model was defined in section 5.3. Here, the parametrization applied for the evaluations in this section is defined. These evaluations use the BS layout with 19 sites and 57 sector cells, which is compliant with 3GPP specifications [3GPP 36.814]. This also implies the application of the matching configuration of the wrap-around geometry model.

The main difference to the previous evaluations is the dynamic data traffic model, which was defined in section 5.3.6. This interacts with the placement of UEs as also described there. The only parameter of the data traffic model is the IAT of the requests. This is a parameter of the following studies. It is normalized to 100 % system load in section 6.2.3.3.

The radio channel and the physical layer are modeled as in the previous evaluations. However, the preprocessed channel traces are used differently here. There are 1000 trace files with a duration of 1 s each. Whenever a UE starts a new transmission, it is assigned a random trace and a random position in that trace. As the simulation continues, consecutive samples are read from the trace. When the trace ends, it is wrapped, i. e. the next sample is the first sample from the trace.

The remaining components of the system model are the same as in the previous section.

#### *6.2.2.2 Evaluated Metrics*

With a full-buffer traffic model, the average UE rate is proportional to the average cell rate and to the system rate. However, this simple relationship does not hold with a dynamic data traffic model. Therefore, the metrics are here separated into cell and user metrics.

The average system or cell rate is not as meaningful in this dynamic traffic evaluations as it is in the previous evaluations. If the system operates under partial load, the cell rate is determined by the offered load. In those situations, delayed data transmissions do not influence the cell rate, but impact experienced network performance of the users. The cell rate is therefore ignorant of performance degradation in these configurations. However, whenever the cell rate falls below the average offered load, this means that the system is incapable of serving all requests. The average cell rate is therefore included in the evaluated metrics. It is complemented with the measured fraction of AC drops.[11]

In the previous evaluations, the average UE rate is used as main metric for network performance.[12] The average UE rate is simple to derive in a full-buffer model. However, its definition is more complex in the dynamic scenario.

Each UE appears in the system only to transmit a single data object. It thus does not make sense to calculate the average rate of a UE over the whole simulated time. Instead, the *experienced rate* is evaluated. This is here defined as the size of a data object in bits divided by the time it took the system to transmit this to the UE in seconds. The duration of the transmission is measured from the arrival of the object at the BS until it is fully reassembled at the UE. Processing times are not modeled. Consequently, the minimal duration is 1 ms. The experienced rate is not measured for messages dropped by the AC mechanism. Analogously to the UE rates in the previous evaluations, the average and the 5th percentile of the experienced rates are evaluated.

One of the objectives of the studies in this section is to evaluate whether the simple prediction mechanism is sufficient to efficiently operate a dynamic system. Therefore, in addition to the cell and user metrics defined in the previous paragraphs, two other metrics are used to evaluate the quality of the prediction. These are, first, the fraction of skipped resource allocations, and, second, the fraction of unused compute resources. Whenever the predicted threshold is too small, that results in overload and the system has to skip the encoding of data for some UEs. In contrast, when the threshold is too low, the utilization of the compute resources is reduced.

#### *6.2.2.3 Reference Configurations*

In the previous studies, the proposed system was compared to an all-encompassing optimum and to a baseline heuristic. The baseline heuristic defined in section 6.1.2.4 is also evaluated in this dynamic model. All-encompassing optimization, however, cannot be performed with the model used in this section. First, dynamic data traffic is difficult to include in the formulation.[13] As it also implicates temporal correlation in the simulation, it requires longer simulation times, which make the optimization more complex. Second, the interference regarded in these studies cannot be modeled consistently in optimization and simulation. This is caused by the mutual interactions of RA, LA, and interference. If variable interference is included in an optimization problem, the optimizer implicitly performs IfCo. However, this cannot be realized optimally in the simulation.

[11] For the definition of the applied AC mechanism see section 5.3.6.

[12] Note that for the full-buffer traffic model applied there, there average UE rate is also proportional to the cell and system rates. However, that is not the case in these evaluations.

[13] Especially, the objective function has to be formulated depending on the experienced rates of the UEs. Proebster et al. [Pro+12] formulated this as binary flags per transmission and subframe, which denote whether a transmission is finished at that subframe. However, that significantly complicates the optimization problem.

Therefore, the all-encompassing optimization problem is not defined and solved for this model. A smaller optimization problem is used instead, which is termed *opt. per subframe* in the evaluations. This problem selects MIMO modes by solving the MCKP introduced in section 5.6.2. It is embedded in the simulator and solved for every subframe. The remainder of this section defines this problem.

Let $r_{u,s,m}$ denote the data capacity of the resources allocated to UE $u$ at subframe $s$ when MIMO mode $m$ is applied. This capacity can be calculated as in the previous problems, i. e. as $r_{u,s,m} = \sum_{n^c \in \mathcal{R}_s^c} a^{(\infty)}_{u,n^c} r_{u,n^c,m}$. However, this reference configuration is realized as component of the simulator, and therefore allows only a single MCS per codeword for all PRBs allocated to a UE. Consequently, the data capacities are calculated by the simulator and do not follow the same definition as those used in the previous problems. Analogously, the processing effort associated with each MIMO mode is calculated by the simulator. This is here denoted as $p_{u,s,m}$. Based on this, the MCKP for a single subframe $s$ is defined as follows.

$$\max_{\substack{\tilde{a}_{u,s,m},\\ u\in\mathcal{U},m\in\mathcal{M}}} \quad \sum_{u\in\mathcal{U}} \mathrm{U}_{\mathrm{cap}}\left(\sum_{m\in\mathcal{M}} \tilde{a}_{u,s,m}\, r_{u,s,m}\right) \tag{6.10a}$$

$$\text{s. t.} \quad \sum_{u\in\mathcal{U}} \sum_{m\in\mathcal{M}} \tilde{a}_{u,s,m}\, p_{u,s,m} \leq P^{\mathrm{peak}}_{|\mathcal{B}|N_{\mathrm{PRB}}} p_{\max} \tag{6.10b}$$

constraints (6.4d) and (6.4g).

#### 6.2.2.4 Resource Allocation Algorithm

The PF heuristic, similar to that introduced in section 3.4.2, is used for RA in the simulations. In each BS $b$, each PRB $f$ is assigned to that UE which maximizes the metric $m^{\mathrm{PF}}_{u,f}$, i. e.

$$a_{u,f} = \begin{cases} 1 & \text{for } u = \arg\max_{u\in\mathcal{U}_b} m^{\mathrm{PF}}_{u,f} \\ 0 & \text{otherwise} \end{cases} \quad \forall u \in \mathcal{U}_b. \tag{6.11}$$

PRBs are assigned successively until each PRB is assigned to a UE or no UEs have data left in their queues. To avoid excessive interference on PRBs with low indices, the PRBs are processed in different random orders for each BS.

First input of the metric calculation is the predicted capacity of the PRBs. In real systems, this is based on CSI reports. Thus, there are errors from noise and quantization. In addition, values received at the BS are in general outdated. Complex mechanisms can be applied at the BS to correct the reports so that robust encoding is possible without unnecessary overhead. This is not modeled in the simulation. Instead, for the calculation of the predicted capacity, which is here denoted as $r_{u,f}(s)$ for UE $u$ on PRB $f$ in subframe $s$, ideal knowledge of all channels is assumed. The channel capacity also depends on the interference caused by neighboring BSs that transmit on the same PRBs. As this depends on the RA of these neighboring BSs, it cannot be idealized without creating circular dependencies. Therefore, for RA purposes it is assumed that all other BSs in the system cause interference. The MIMO mode is selected to maximize the data capacity, i. e. potentially limited processing resources are not taken into account.[14]

The data capacity of a PRB can be calculated individually by performing LA for that PRB, i. e. selecting the best MIMO mode and MCS. However, this does not reflect the capacity which can be realized when the PRB is assigned to a UE and encoded together with other PRBs. That is caused by potentially different MIMO modes and MCSs being optimal for the other PRBs, which enforces the BS to make a compromise in the final LA. To avoid this error in rate prediction, in the simulations the calculation of $r_{u,f}(s)$ takes the previously assigned PRBs into account. Whenever $r_{u,f}(s)$ is calculated, LA is performed for the combination of the already assigned PRBs and the current PRB $f$. The value of $r_{u,f}(s)$ is then the difference between the capacity of the previous allocation alone and the capacity of the combination.

Second input of the metric calculation in equation (3.11) is the average of the rate allocated in the previous subframes. Equation (3.12) describes the calculation for the general case. There, the average rate $\bar{R}_u(s)$ is calculated by filtering the actually allocated rates $r_u^{\text{act}}(s)$ with a low-pass filter. To avoid unwanted interactions with MIMO mode selection, the history is here updated at each subframe with the sum of the predicted capacity of the allocated PRBs. So, when $a_{u,f}$ denotes allocation of PRB $f$ to UE $u$ in the current subframe $s$, the allocated rate is calculated as $r_u^{\text{act}}(s) = \sum f \in \mathcal{F} a_{u,f} r_{u,f}(s)$. It is thus neither influenced by the adapted MIMO mode selection nor by potentially canceled encoding.

When a new data transmission starts, i. e. a new UE enters the system, no resources have been allocated to this UE before. Therefore, the average of the previously allocated rates is zero. When strictly following equation (3.11) for calculation of the allocation metric, this results in $m_{u,f}^{\text{PF}} = \infty$. Consequently, all PRBs of the current subframe would be allocated to that UE, before the average is updated for the first time and the metric takes a reasonable value. This undesired behavior is counteracted by a modification of the metric calculation. In the simulations, the metric for allocation of PRB $f$ to UE $u$ is calculated by

$$m_{u,f}^{\text{PF}} = \frac{r_{u,f}(s)}{\left(1 - \frac{1}{t_c}\right) \bar{R}_u(s-1) + \left(\frac{1}{t_c}\right) \sum_{f \in \mathcal{F}} a_{u,f} r_{u,f}(s)}. \tag{6.12}$$

where $a_{u,f} = 0$ is assumed for those PRBs which are not assigned to any UE yet. So, instead of dividing by the average rate updated after allocation of the previous subframe $s-1$ alone, the already performed allocations for the current subframe $s$ are taken into account. This anticipates the update of the average after the current subframe $s$. With this modification, $m_{u,f}^{\text{PF}} = \infty$ only for the first considered PRB.

To further quicken the convergence of $\bar{R}_u(s)$ to the stationary average, the value of $t_c$ is adapted over time. Let $s_u^{\text{start}}$ denote the subframe when the transmission to UE $u$ started, i. e. that subframe where it gets the first PRB assigned. The value of $t_c$ for this UE at subframe $s$ is then defined as

$$t_{c,u}(s) = \min\left(s - s_u^{\text{start}} + 1, t_c^{\max}\right) . \tag{6.13}$$

[14] Note that these simplifications are only made for the RA algorithm. After RA has been performed by all BSs, the real interference occurring in that subframe is known. That is then used as the basis for LA, where the proposed system or the baseline heuristic are used to select the MIMO modes actually used for transmission.

This gives additional weight to the current values in the subframes shortly after the new transmission arrived in the system. The weight gradually decreases and approaches the weight corresponding to the maximum time constant $t_c^{\max}$. This parameter is here configured as $t_c^{\max} = 1000$.

The described RA algorithm, which includes the repeatedly performed LA, contributes significantly to the simulation complexity. The complexity scales linearly with the considered candidate UEs. To reduce it, the described heuristic is prepended with a simple filter, which limits the considered candidate UEs per subframe and BS to $N_{\text{cand}} = 32$.[15] These are selected in a round-robin fashion from all active UEs served by that BS.

In a preliminary evaluation, this RA heuristic is compared to the optimizer in the same scenario as evaluated in section 6.1. There, it achieves about 99 % of average UE rate realized by the optimizer. The same is true for the 5th percentile of the UE rates.

#### *6.2.2.5 Implementation of the Simulation Model*

Small-scale effects of the radio channel and the offline component of the PHY layer model are calculated in *MATLAB*. There, channel trace files are generated, which are then read by the main simulator. That is based on the *IKR SimLib* [IKRSimLib] and *IKR RadioLib* and is mainly implemented in *Java*. Some performance-critical components (e. g. the RA algorithm) are implemented in *OpenCL*. These use the *OpenCL Runtime for Intel Core and Intel Xeon Processors* for efficient parallelization on standard *central processing units* (CPUs). For the reference configuration *opt. per subframe*, the *IBM ILOG CPLEX Optimizer* is called repeatedly to optimize the MCKP for each subframe.

#### *6.2.2.6 Procedure of Simulation*

In contrast to the previous evaluations, which are conducted in independent drops, the simulations for this section use the batch-means method and therefore consist of a single continuous run for each parametrization [Law07]. This is reasoned as follows.

In the full-buffer traffic model, the positions of the UEs are fixed. Therefore, independent drops are required there to sample different constellations. However, with the dynamic traffic model UEs are repeatedly placed on new positions. Thus, the argument for independent replications does not hold any more. At the same time, the dynamic traffic model implicates that the system requires a significant warm-up time, where its load is not stationary but still influenced by the initial conditions.[16] The batch-means method allows to have only a single warm-up phase for each parametrization. This is therefore applied here to reduce the computational effort for the simulations.

[15] Gains from opportunistic RA do not improve further when considering more UEs. E. g., Ellenbeck et al. [Ell+09] show that there is no significant difference between considering 20 or 40 UEs. A power of 2 has been chosen here because of implementation reasons.

[16] It is difficult to reliably determine startup conditions which allow to initialize the system in a stationary state. Therefore, the simulations are started without any UEs transmitting. As waiting UEs slowly accumulate in the system, that requires a significant amount of time to reach a stationary state.

The duration of the warm-up phase is configured to be $t_{\text{warm-up}} = 300\,\text{s}$. This value is derived from pilot runs in section 6.2.3.1. Configurations where this is not sufficient to reach a stationary system state are detected by statistical tests, which is described in section 6.2.2.7. During this warm-up phase, no metrics are evaluated.

To be able to calculate confidence intervals for the evaluated metrics, the simulated time following the warm-up phase is divided into $N_{\text{batches}} = 10$ batches. Each batch has the duration $t_{\text{batch}} = 60\,\text{s}$. This duration is derived from pilot runs in section 6.2.3. The evaluated metrics are first averaged over all events in a batch.[17] The batch-averages are then used to calculate the 95 % confidence interval as described in section 5.5.1.

#### *6.2.2.7 Detection of Non-Steady Behavior*

The system can operate in one of three stationary states. In case the offered load is low compared to the capacity of the system, the RA algorithm does not allocate all PRBs. Therefore, the interference is low, which results in high spectral efficiency. Active UEs do not accumulate, because they are served more quickly than new UEs arrive. When the load of the system is sufficient to utilize all PRBs in all cells, transmissions take more time to complete. That results in an increasing number of UEs in the system, because arrivals still occur at the same rate. However, the more UEs are waiting, the more flexibility the RA algorithm has to assign resources. Thus, the spectral efficiency increases. This counteracts the increasing load, and can lead to a balanced state. Finally, in case of extreme overload, the number of UEs is limited by the AC mechanism.

It is plausible that slow transient behavior occurs when the system is loaded only marginally above its maximum capacity. In that case, it takes a long time until so many UEs have accumulated that they are limited by AC. Thus, it is not possible to give an upper limit for the required warm-up time. These configurations cannot be avoided, because the system capacity also depends on the amount of available compute resources, which is a parameter of the simulations. Therefore, a hypothesis test for trends is used to detect these configurations.

The experienced rates of the data traffic objects are used as metric to which the test is applied.[18] These are evaluated in the order in which the objects are completed, ignoring objects which are completed during the transient phase. The variance of the experienced rates is high, which impairs the performance of the test. Therefore, the samples are split into 100 equally sized groups, so that each group contains consecutive samples.[19] Subsequently, the average is calculated for each group. The unweighted test of Cox and Stuart [CS55], denoted as $S_3$ in their publication, is then applied to the sequence of averages.

The null hypothesis of the test is that there is no trend in the samples. As the number of group-averages is large, the result of the test is assumed to be normally distributed. The CDF of

[17] So, e. g., the experienced rate is calculated for each object for which the transmission is completed. Then, the average is calculated over all objects which complete during a batch.

[18] Other metrics could also be used. One obvious approach is to count the actual number of UEs in the system. Another one is to sum up the total data waiting to be transmitted in all queues (see section 6.2.3.1). The experienced rate is chosen here because it is also one of the most important evaluated metrics.

[19] This approach is similar to the batching described below. However, more groups of samples are used here, because the higher number has shown to give more sensible results for this test.

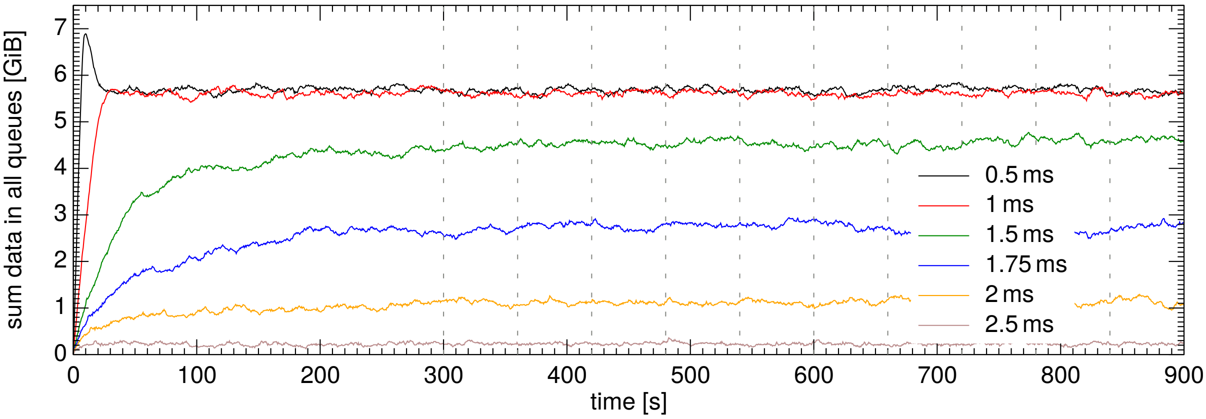


**Figure 6.3:** Determination of the warm-up phase. Each curve corresponds to a different average IAT.

the normal distribution is used to derive the probability that the result of the test occurs although the null hypothesis holds. It is here assumed that a trend arises in that simulation run whenever this probability falls below 5 %. The respective parametrizations are marked in the plots.

### 6.2.3 Configuration and Calibration of the System

The procedure of simulation described in section 6.2.2.6 requires configuration of the durations of the warm-up phase and the batches. This is performed by evaluating a set of pilot runs in sections 6.2.3.1 and 6.2.3.2, respectively. The same pilot runs are used to derive the capacity of the system in section 6.2.3.3. Based on the capacity, the offered load to be used in the following studies is defined.

For the simulations in this section, the evaluation methodology is the same as for the following studies. However, the compute resources are not restricted. Consequently, none of the compared reference configurations applies here.[20]

#### *6.2.3.1 Configuration of the Duration of the Warm-Up Phase*

Two effects occur during the transient phase. First, small objects and those for UEs with high channel capacity are transmitted quickly, while objects which take more time to transmit accumulate in the system. Second, the number of active UEs increases, until either the efficiency of the opportunistic scheduling suffices to serve the offered load or the AC limit is reached. As a metric for these effects, the sum of the data waiting to be transmitted in all queues is evaluated.[21]

This metric is plotted over the simulated time in figure 6.3 for different average IATs. The plot shows that the data in the queues accumulates until it reaches a load-dependent value. In case of extreme overload, the AC mechanism and the distribution of object sizes restrict the total amount

[20] Note that the proposed system and the reference configurations only influence the behavior of the system in configurations where compute resources are limited.

[21] The experienced data rate, which is used in the statistical tests described in section 6.2.2.7, can also serve to derive the duration of the warm-up phase. However, transient effects have shown to be less clearly visible from a plot of that metric.

of data in the queues.[22] From the plot can be derived that for the pilot simulations performed here, $t_{\text{warm-up}} = 300\,\text{s}$ seems to be a reasonable configuration.

#### *6.2.3.2 Configuration of the Duration of the Batches*

After cutting off the warm-up phase, the remainder of the simulated time is divided into equally sized batches. Their size has to be chosen such that there are a sufficient number of samples per batch and such that the correlation between consecutive batches can be neglected.

The evaluated metrics are based either on a sample per data object or on samples occurring per subframe. As the IAT $t_{\text{IAT}} = 1$ ms already constitutes an overloaded system, samples associated with data traffic objects occur less frequently than the other types of samples. However, even an IAT of $t_{\text{IAT}} = 10\,\text{ms}$ still results in 100 data objects per second, so the number of events is considered to be not the limiting factor for the batch duration.

Long-term correlation in the model used here can be caused by large data objects that are transmitted to UEs which have low capacity channels.[23] To estimate the duration of these effects, the distribution of transfer times of data objects is evaluated (no plot shown). The outcome of that evaluation is that in general, the transfer of data objects takes more time the higher the load of the system is. Under overload (e. g. $t_{\text{IAT}} = 0.5\,\text{ms}$), 90 % of the objects are transmitted in less than 32.5 s, and 98 % finish transmission in 55.5 s.[24]

As result of this evaluation, the batch duration is set to $t_{\text{batch}} = 60\,\text{s}$. While this can still cause some correlation of consecutive batches, no effects are expected that influence the next but one batch. This duration is chosen as a compromise to keep simulation complexity at a reasonable level. The total simulated time is $t_{\text{total}} = t_{\text{warm-up}} + N_{\text{batches}} \cdot t_{\text{batch}} = 900\,\text{s}$ for each configuration.

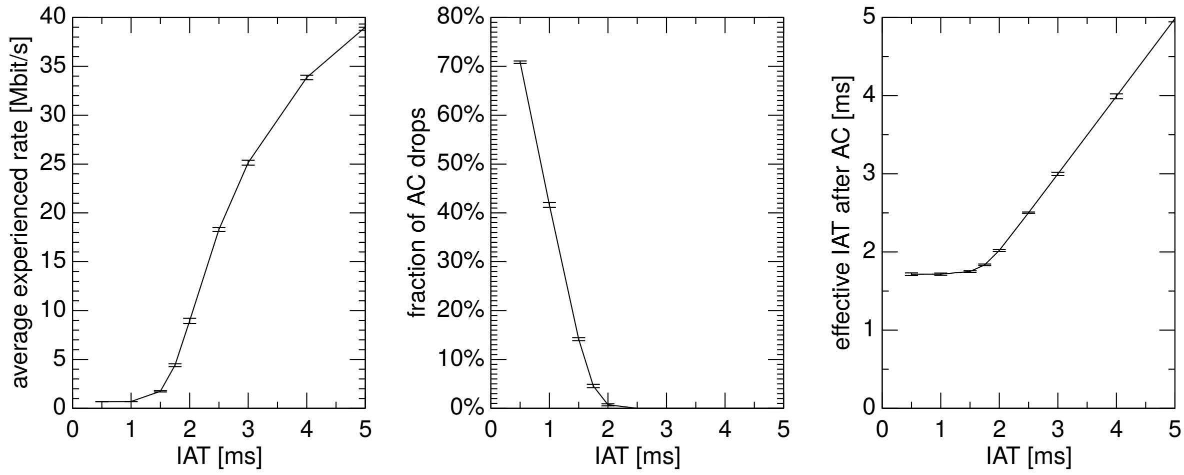
#### *6.2.3.3 Calibration of the System Load*

Besides the parametrization of the warm-up phase and batch duration, the preliminary simulations are also used to normalize the system load. The higher the system load is, the lower is the experienced rate. When the system load increases above a certain threshold, some requests get dropped by the AC mechanism. The *effective IAT* is introduced here as a metric for the carried load. It is defined as the average time between two consecutive requests that pass the AC mechanism. In a stationary system state, this is equivalent to the average time between two consecutive completed transmissions. These three metrics are shown in figure 6.4.

[22] The maximum size of traffic objects is 108 MiB. AC limits the number of active UEs to 5700. Therefore, the absolute maximum amount of data in the queues is 601 GiB. The stationary amount of data accumulated in the case of overload is, however, difficult to determine analytically. It is determined by the effect that small objects are quickly transmitted and replaced by new objects, while large objects stay in the system for longer time.

[23] E. g., in a configuration with low load, such a data object can cause a single BS to use all PRBs for a significant amount of time. This causes interference to other cells and increases compute resource utilization. The duration of such an effect is extended if multiple large objects happen to be transmitted simultaneously by the same BS.

[24] Note that, when the system capacity decreases because of limited compute resources, the transmissions also take longer. However, for reasonable configurations, i. e., those for which the experienced rate is plotted in figure 6.6, never more than 0.5 % of the transmissions take more than 60 s.



**(a)** Average experienced rate **(b)** Drops performed by AC **(c)** Effective IAT

**Figure 6.4:** Calibration of the dynamic traffic model

From figure 6.4c can be derived that the effective IAT does not fall below 1.72 ms. This is considered to be the capacity of the system. It is, however, not expected that the system is operated at this load, because there AC drops already occur and the experienced performance is low. For the following studies, fractions of 80 %, 60 %, and 40 % of this load are evaluated. This corresponds to $t_{\mathrm{IAT}} = 2.15$ ms, $t_{\mathrm{IAT}} = 2.86$ ms, and $t_{\mathrm{IAT}} = 4.29$ ms, respectively. With the average object size of 688 KiB, this results in data rates of 46.1 Mbit/s, 34.6 Mbit/s, and 23.0 Mbit/s per cell.

## 6.2.4 Studies

Simulations have been performed based on the system model and the calibration from the previous sections. The results of these simulations are presented in this section. It is structured according to the metrics introduced in section 6.2.2.2. First, section 6.2.4.1 evaluates metrics on cell level, which show under which conditions the systems are capable of serving the offered load. Second, section 6.2.4.2 focuses on the network performance as seen from the perspective of the users. Finally, section 6.2.4.3 evaluates additional metrics to assess the performance of the prediction component of the proposed system.

### *6.2.4.1 Metrics on Cell Level*

The average transmitted data rate per cell is plotted on the $y$-axis of figure 6.5a. The $x$-axis of that figure shows the compute resource limit. The curves in the plot represent the three configurations *opt. per subframe*, *proposed*, and *baseline*, each of which is evaluated with 40 %, 60 %, and 80 % offered load. The former are differentiated by colors, the latter by line styles.[25]

For each evaluated load, the compute resources required for unconstrained operation are marked by a vertical bar on the respective curve.[26] In addition, for each parametrization, the lowest

[25] Solid, dashed, and dotted lines denote 40 %, 60 %, and 80 % load.

[26] Strictly, the horizontal position of the bar is equivalent to the compute resources which are sufficient for unrestricted operation in 99.9 % of the subframes, given that the available compute resources are not limited. The

compute resource limit where more than 99 % of the offered load is served is marked with ×. Parametrizations which show non-steady behavior are not marked here, because it is assumed that the cell rate does not increase significantly when further UEs accumulate.

For large amounts of compute resources, the average cell rate is determined by the offered load. When the compute resources are restricted, the cell rate starts to drop at different values. In general, with higher offered load the system requires more compute resources to serve this load.

Compared to the compute resource usage in the unconstrained case, the configuration *baseline* requires between 7 % and 14 % less compute resources to serve the offered load.[27] The required resources are roughly equal to the average compute resource utilization in the unconstrained system. This can be explained by an averaging effect in time: When UEs are skipped in the encoding module, their transmissions are retried at a later subframe. This delays the data objects. However, as long as the compute resources are not constantly exhausted, the system can complete all requests and waiting requests do not accumulate. Thereby, peaks in the compute resource utilization are cut off and shifted to later subframes.

In addition to this shifting effect, the proposed system also increases the efficiency of the compute resource usage, i. e. it can transmit more bits per compute effort. This allows to maintain the full cell rate with a significantly lower amount of compute resources. Thus, for 40 % load, 10 % compute resources are sufficient, which is a quarter of the peak compute resources utilization in the unconstrained case.[28] Similarly, for 60 % and 80 % load, 20 % and 35 % compute resources are required. This roughly corresponds to a third and half of the peak resource utilization, respectively. The same is true for *opt. per subframe*, i. e. there is no significant difference between the cell rates achieved by the proposed system and the optimization per subframe.

When the compute resources are limited to values below the marked thresholds, the cell rate drops. The curves for different loads approach a common limit, which can be interpreted as the capacity of the system depending on the compute resource limit. This capacity seems to be a linear function of the compute resources for *baseline* (about 850 kbit/s per 1 % compute resources). For *opt. per subframe* and *proposed*, a linear dependency cannot be identified clearly.

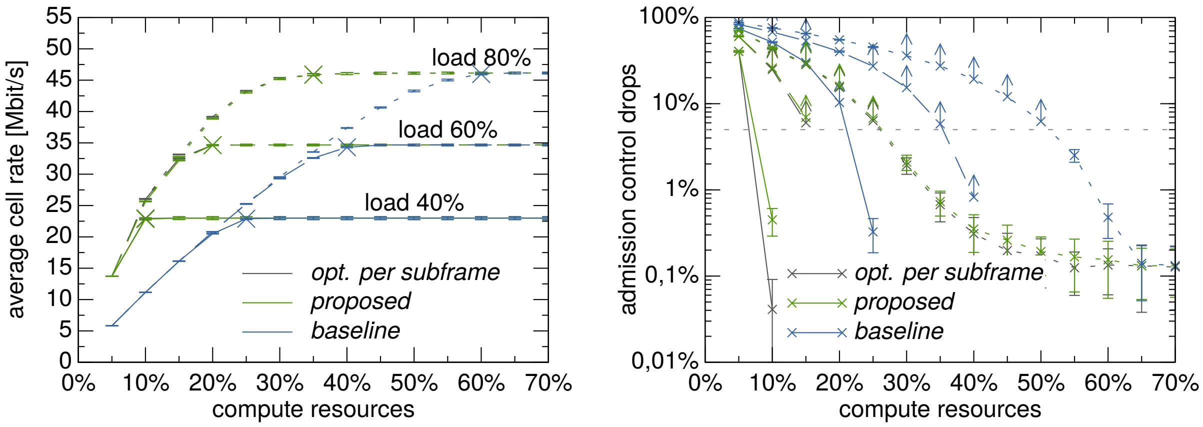
When the average cell rate drops below the offered load, that means that the system cannot serve all requests, but some are dropped by the AC mechanism. This is shown in figure 6.5b. Here, the fraction of messages dropped by AC is plotted over the compute resource limit. The $y$-axis is scaled logarithmically, so that differences between small as well as between large drop rates are visible. Points with zero drop rate are omitted from this plot. For parametrizations for which a trend is detected as described in section 6.2.2.7, the confidence interval is replaced with an arrow pointing upwards. Colors and line styles are the same as in figure 6.5a.

Figure 6.5b shows that for 80 % load, there are about 0.1 % to 0.2 % AC drops even with 70 % compute resources. Here, the variations of the data traffic cause single cells to be in temporary overload. The fraction of dropped requests rises as the compute resources are limited. For *baseline*, 10 % of the requests are dropped when only 50 % compute resources are available.

vertical position of the bar does not carry any meaning, but is adjusted so that the bar lies on the curve of the respective load.

[27] This is seen in the plot as the horizontal distance between the respective markings.

[28] The resource utilization in the unconstrained case is defined as described in footnote [26].



**(a)** average rate per cell

**(b)** fraction of messages dropped by AC

**Figure 6.5:** Evaluation of cell rate and AC drops

With the same limit, *opt. per subframe* and *proposed* drop only 1.7 % and 1.9 % of the requests, respectively. For lower system loads, the drop rates increase at lower compute resource limits. This matches the results from figure 6.5a.

It is here assumed that the system will not be operated at a drop rate of more than 5 %. It can be argued that the users of the network will not accept a service which fails to serve 5 % of the requests. This threshold is marked with a horizontal dashed line in figure 6.5b. For parametrizations with higher drop ratios, the following section does not evaluate the user level metrics.

#### *6.2.4.2 Metrics on User Level*

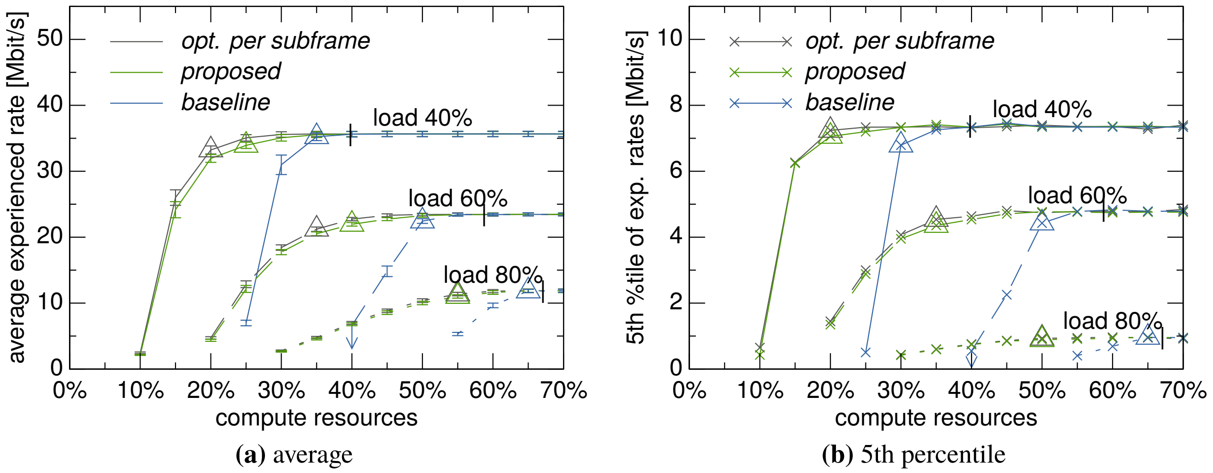
The previous subsection discussed at which configurations the system is capable of serving all requests. It did, however, not consider the performance perceived by those users who get their data objects transmitted. This is the focus of this section.

Figures 6.6a and 6.6b show the average and the 5th percentile of the experienced rates, respectively. The plots show the same set of configurations as the previous plots. As before, the compute resources required for unconstrained operation are marked by a vertical bar.[29] In addition, in both plots the compute resources sufficient to maintain 90 % of the unconstrained experienced rate are marked with a triangle. The only parametrization for which a trend is detected as described in section 6.2.2.7 is designated with an arrow pointing downwards. This symbolizes that the stationary experienced rate is probably lower than the plotted value. As stated in section 6.2.4.1, points where the AC drops more than 5 % of the requests are omitted from the plot.

In general, the experienced rate is higher when the offered load is lower, because in that case less UEs compete for resources and the average level of interference is lower. Compared to the cell rate in figure 6.5a, the experienced rates drop earlier. The curves for the average and the percentile follow the same shape, so the cell border users are neither preferred nor penalized.

It is reasonable that for all configurations the experienced rate drops as soon as the compute resource limit falls below the resources required for unconstrained operation. Independently of

[29] See note [26] on page 171.



**(a)** average **(b)** 5th percentile

**Figure 6.6:** Evaluation of experienced UE rates

whether the system copes with the limited compute resources by only adapting MIMO modes or by skipping to encode some PRBs, the transmissions are not served as quickly as without limit. However, for *baseline*, the experienced rates degrade significantly when reducing the compute resources to values more than 5 % below the marked values. At the same compute resource limit, the degradation is barely visible for the configurations *opt. per subframe* and *proposed*.

To maintain experienced rates equivalent to 90 % of those achieved with unlimited compute resources, the configuration *baseline* requires about 10 to 15 percentage points (or 30 % to 50 %) more compute resources than the remaining configurations. Even when the compute resources are limited further, the performance achieved with the configurations *opt. per subframe* and *proposed* degrades more smoothly than that achieved by *baseline*.

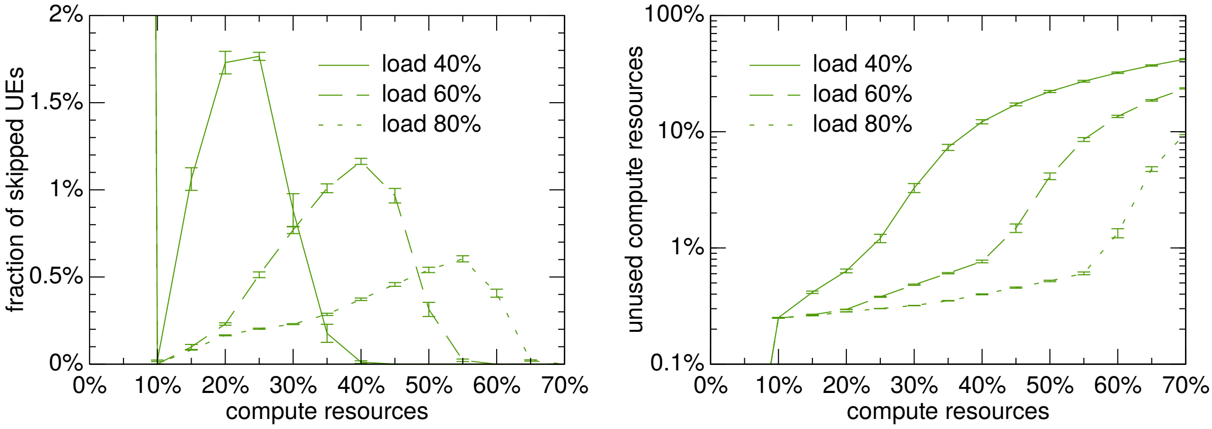
#### *6.2.4.3 Metrics for the Performance of the Prediction Mechanism*

The evaluations presented by the previous sections have shown that the performance achieved by the proposed system is close to that achieved with optimal MIMO mode selection. To maintain the same performance as those two configurations, the baseline system requires significantly more compute resources. This section complements these results by an evaluation of two internal metrics of the system.

Figure 6.7a plots the fraction of skipped sets of PRBs over the compute resource limit. Adaptation of MIMO modes alone cannot reduce the processing effort to values below 10 %. Therefore, many UEs are skipped for $p_{\text{max}} = 5\,\%$. For higher compute resource limits, the fraction of skipped UEs does not rise above 2 %.

As expected, no UEs are skipped when the compute resources are effectively unlimited. When lowering the limit, the fraction of skipped UEs rises to a load-dependent maximum and then falls again. The general shape can be explained by the variance of the compute load, which is higher when the system does not use all PRBs. The most UEs are skipped for 40 % load, and the peak fraction is lowered as the load increases.

The amount of unused compute resources is shown in figure 6.7b. The $y$-axis is shown in logarithmic scale, so that small values can be differentiated in the chart. In case more compute



**(a)** fraction of skipped UEs

**(b)** unutilized compute resources

**Figure 6.7:** Evaluation of internal metrics

resources than required are available, low utilization is expected and acceptable. When the compute resources are limited, the amount of unused compute capacity quickly falls below 1 %.

The three curves show bends at about 20 %, 40 %, and 55 % compute resources for 40 %, 60 %, and 80 % load, respectively. At these points, the reduced spectral efficiency of the low-complexity MIMO modes forces the system to use all PRBs. At lower compute resource limits, under-utilization of the compute resources is directly equivalent to a loss in system performance. The higher the load, the more quickly the under-utilization decreases when the compute resource limit is lowered. This matches the outcome of figure 6.7a. It can be interpreted as the proposed system performing better at higher load.

When reducing the compute capacity down to 10 %, the under-utilization approaches the value 0.25 % for all configurations. This equals the configured value of the offset parameter $p_{\mathrm{off}}$. For 5 % compute resources, only the simplest MIMO modes are selected and the system performance is determined by a large fraction of skipped UEs. Thus, the under-utilization of compute resources approaches zero.

The compute resource utilization of the configuration *baseline*, which is not shown in the plot, is higher. There, the under-utilization approaches 0.04 % for low compute resource limits. This value is determined by the granularity of compute jobs. Due to the fact that data for a UE is either encoded completely or skipped as a whole, in average half of the compute effort of such a compute job cannot be utilized. The utilization achieved by *opt. per subframe* is even higher, because there the optimizer can combine MIMO modes to make best use of the available capacity. However, the achieved network performance of the two reference configurations differs significantly. Thus, high utilization of the compute resources does not guarantee high performance.

### 6.2.5 Summary and Conclusion

The main objective for the evaluations in this section is to show whether a simple prediction mechanism is sufficient to handle dynamic load. An additional objective is to demonstrate the effect of partial system load on the compute resource requirements. Dynamic traffic and interference models are applied, because it is assumed that these effects are the main contributors to variations of the compute load.

Section 6.2.4.1 evaluated metrics on a cell level. These allowed to derive the amount of compute resources which is required to serve the offered load. By shifting the compute effort in time, the baseline heuristic can cope with a compute resource limit roughly equal to the average compute resource utilization in an unconstrained system. Compared to the peak resource utilization, this saves between 7 % and 14 %. At the same time, the proposed system uses the resources more efficiently. That allows to serve the offered load with a quarter to half of the peak utilization. As expected, most resources are required for a highly loaded system, because there the capacity headroom is lowest. A marginal difference between the proposed system and the optimization per subframe is only visible in the fraction of messages dropped by AC.

The performance from the perspective of the users was studied in section 6.2.4.2. When limiting the compute capacity, the experienced rate drops much earlier than the cell rate. For a certain range of compute resource limits, the system can serve the offered load by using less complex MIMO modes or by skipping the encoding of transmissions. However, both delays the transmissions of the data objects. By using the compute resources more efficiently, the proposed system can maintain the same experienced rate as the baseline heuristic with about 10 % to 15 % less compute resources.

Section 6.2.4.3 studied internal metrics of the system to evaluate the performance of the applied prediction mechanism. The proposed mechanism achieves a low fraction skipped encodings, which is typically below 1 % and never higher than 2 %. From this, no significant negative impact on the remaining components of the RAN is expected, because these numbers are much lower than the typical decode error rate. The system achieves a high utilization of the available compute resources. The simple control loop, which is used in the proposed system for the prediction of the efficiency threshold, already provides reasonable results. However, both metrics might be improved by a more sophisticated control loop or other prediction mechanism.

This section has shown that the proposed system can cope well with the dynamics caused by realistic data traffic and interference models. The simple control loop, which predicts the efficiency threshold based on the compute load in previous subframes, already leads to results which are close to the optimum. The variations of the load over time allow to cope with moderate compute resource shortages by just delaying transmissions. However, by using the resources more efficiently, the proposed system can maintain nearly unimpaired network performance with significantly lower compute resource limits.

# 7 Summary and Conclusion

## 7.1 Summary

Subject of this thesis is the design of an efficient mechanism which achieves elastic utilization of compute resources in a cellular communication system. This mechanism is developed and evaluated using the example of an LTE system. Thus, chapter 2 introduced LTE and its architecture. Originally, the LTE RAN is designed as a large number of self-contained eNodeBs distributed in the field. This is modified by the concept of C-RAN, which centralizes the signal processing of multiple eNodeBs in a central pool of BBUs.

LTE is optimized to make efficient use of the radio channel. OFDM splits a large bandwidth up into narrow subcarriers. OFDMA allows to assign these (on a coarser granularity of PRB pairs) to different users. This is used by a flexible RA, so that each user is served on those parts of the radio channel where a high capacity can be achieved. For each user, LA configures modulation and coding to match the actual capacity of the assigned resources. In addition, it selectively applies MIMO to transmit multiple spatially separated streams in parallel.

This is all controlled by the eNodeB. Thus, all decisions have to be communicated to the respective UEs to enable these to decode the transmissions. The flexibility is partially restricted by the signaling capabilities of the eNodeB. These are standardized as part of the air interface specification. RA and LA are not part of the standardization, but can be implemented individually by eNodeB vendors.

Typically, LA tunes modulation, coding, and MIMO to make maximum use of a set of allocated resources. This influences only the single user that is receiving or transmitting the data. In contrast, RA is the central mechanism in LTE that balances between the demands of different users. Each user may apply different metrics for QoS, e. g., throughput, packet drop rates, and packet latency. When handling competing demands, RA mechanism can either try to achieve fairness or prefer single users, e. g., those with favorable channel conditions. Fairness often comes at the cost of reduced average performance. A widely used fairness definition is that of proportional fair. It can either be formulated as objective of an optimization problem, or be realized by a matching heuristic. More advanced RA heuristics do typically also take other parameters such as queue length or differing requirements into account. Requirements, optimization problems, and heuristics for RA were discussed in chapter 3.

While RA copes with competing demands inside a cell, neighboring cells do also influence each other. To make best use of the available bandwidth, LTE is designed so that the same spectrum can be used in adjacent cells. However, that results in high interference for those UEs which are located close to the border of a cell's serving area. There are multiple approaches to handle this. Either receivers can be enabled to suppress interference, or transmitters can coordinate to reduce the interference. This IfCo can be seen as a RA problem spanning multiple cells. So, analogously to a normal RA problem, it can be approached by optimization or with heuristics. Chapter 4 gives an overview over all related aspects. It especially focuses on the optimization, because IfCo is also part of the optimization problem used as foundation for the design of the proposed mechanism. Different sources of literature are compared by unifying their understanding of coordinated resources to a common resource model. In addition to mere coordination, neighboring eNodeBs can also collaborate more closely to simultaneously serve their users. This is, however, not focus of this thesis.

Signal processing required to encode data that is to be transmitted and to decode received data generates processing effort. This effort contains a rather static share, which is related to cell specific functions (e. g. reference signals and OFDM processing). In addition, it contains a more dynamic share, which is associated with individual users and therefore influenced by RA and LA. This thesis focuses on the dynamic share occurring during DL transmission of data.

Varying demands of users, varying radio channels, and dynamically changing RA cause fluctuations of the total compute effort. These can be partially limited by combining the processing of multiple cells in a single central BBU pool. However, to efficiently cope with these fluctuations, a mechanism is required which allows to cut the peaks without significantly impacting the network performance. The same mechanism can also facilitate the implementation of a BBU as software running on a non-RT system. Having an efficient way to achieve elastic utilization of compute resources allows to forgo without stringent planning and management of these resources. This is beneficial, because these often come with high complexity and low resource utilization.

This thesis presents such a mechanism, together with the reasoning leading to its design and a thorough performance evaluation. The proposed mechanism makes the compute resource utilization of a mobile communication system elastic, i. e., it allows the computational complexity to dynamically adapt to the available resources. In doing so, it maintains high network performance under all but the most stringent resource limits.

Mobile communication systems provide multiple ways to trade network performance off for reduced computational complexity. The related decisions are, however, typically made at design time of a system, e. g., by selecting certain algorithms or implementing them with a certain numerical accuracy. Proposals to perform dynamic adaptation, which are available in literature, focus on the requirements of single links and do not deal with the complexity of a whole system. Approaches related to this thesis also exist in the subject of RT scheduling. However, these do not capture the special requirements of mobile communication systems. Thus, they either rely on more flexible system models or on a central coordinator. The processing resources for mobile communication systems are typically planned statically. Some publications also propose a dynamic resource management. They do, however, either only target the allocations of functions to compute units, or are content with simply dropping transmissions in case the resources are overloaded.

The simulations and optimizations applied in this thesis use a common system model defined in section 5.3. The model for the radio network mainly follows the guidelines specified for system level simulations by 3GPP. It gives a realistic estimation of the channel capacity achieved by different MIMO modes. This is complemented by a model for the processing effort, which also takes the LA parameters into account. Three variants of this system model are used in the different evaluations.

In section 5.5 a small scenario with 21 cells was used to limit the complexity. In addition, a full-buffer traffic model was applied and variations of the radio channel over time and frequency were neglected. Section 6.1 also used the small scenario and the full-buffer traffic model. It did, however, consider the variation of the radio channel over 10 MHz bandwidth and 100 ms evaluated duration. To compensate for this additional complexity, IfCo was not modeled there. Section 6.2 finally did not consider an all-encompassing optimization problem. This allowed to use the larger scenario with 57 cells and model IfCo as well as dynamic data traffic. In addition, the evaluated time frame was extended to match the dynamic variations of the data traffic.

The design of the proposed adaptation mechanism is based on insights derived from solving an optimization problem. This problem was defined in section 5.4, based on the resource model introduced for the comparison of IfCo problems in section 4.2. It jointly evaluates IfCo, RA, and MIMO mode selection. To achieve a manageable problem size, only the interference caused by the strongest three interferers is modeled. Different variants of the problem are defined to model different fairness requirements. Orthogonal to that, other variants restrict the problem so that only subsets of the variables are adapted to the available processing resources.

Solutions to these optimization problems were then studied in section 5.5. The evaluations show that the required processing resources depend on the desired fairness scheme. The highest throughput is achieved when fairness is not considered. In contrast, fair systems achieve a lower throughput, but also require less processing resources. This is caused by differences in selected MIMO modes and in the fraction of resources which remain unused to reduce interference. When jointly adapting all variables, the system can cope well with limited processing resources. It achieves 90 % of the original throughput with only 30 % to 35 % of the peak processing resources. This is realized by simultaneously changing MIMO modes and increasing the fraction of empty resources. When only adapting subsets of the variables, the performance is reduced. The least significant drop is achieved by adapting MIMO modes individually for each UE.

The proposed mechanism is designed based on these findings. It adapts the MIMO modes used to serve the UEs to the channel conditions and available processing resources. The selection of the modes is interpreted as MCKP. A distributed heuristic is developed which is based on a known greedy solving heuristic for such problems. It consists of three components. A MIMO mode selection algorithm is executed individually for each UE. It selects a mode based on the UE's channel conditions and a global threshold value. A fallback mechanism ensures stable system operation in case the selection algorithm chooses too complex MIMO modes. Finally, a prediction mechanism defines the value of the global threshold based on the overload experienced in previous subframes.

This mechanism can be integrated into a BBU without modifying existing components performing IfCo and RA. It introduces only limited additional complexity, because the selection algorithm, which resembles the most complex part of the proposed mechanism, can be executed independently for each UE. It is also robust, because it can tolerate variations in the available processing capacity as well as deviations in the estimated complexity of the MIMO modes.

The evaluation of the proposed mechanism is split into two parts. First, its performance was compared to that of differently constrained optimization problems in section 6.1. The studies show that the proposed system achieves performance comparable to the optimizer. Compared to the all-encompassing optimization problem, the UE rates drop by 2 % to 11 %. Here, the smaller deviations are achieved in situations with moderate compute resource overload.

The proposed mechanism relies on a component which predicts the efficiency threshold based on previous subframes. It thus depends on a temporal correlation of the compute load. In section 6.2, it was checked whether high performance can be maintained in a dynamic scenario. Those evaluations show that the proposed system achieves significant gains over a simple baseline mechanism. Its performance is close to that achieved by centrally optimizing the MIMO mode selection at every subframe.

## 7.2 Conclusions

The proposed system has shown to be successful in adapting the compute requirements of a mobile communication system to the available resources. While achieving this elasticity, it also maintains a high efficiency, which is close to that of an optimal solution. During the design of the mechanism, an encompassing view of the whole LTE system was kept in mind. In addition, realization aspects have been considered. Consequently, the proposed mechanism is ready to be integrated into the LTE system.

The proposed system allows to cope with moderate compute resource overload without noticeable impact for the users. Depending on the expected data traffic, only 25 % to 55 % of the compute resources required to handle the theoretic peak situation are sufficient. These resource savings can be realized without noticeably impacting the network performance perceived by the users. The mechanism thus facilitates a tight and economical dimensioning of compute resources.

Furthermore, the mechanism also allows for a more dynamic resource management, which is typical in IT cloud environments. With elastic resource utilization, the system can tolerate variations in the available compute power. These can be caused, e. g., by the OS or by other virtual instances running on the same hardware. It also facilitates a network operator to switch off unused hardware units in low load times. When then the load increases again, the proposed mechanism can bridge the time until the hardware is brought online. Similarly, it maintains stable network operation in case of hardware failures.

In this thesis, the design is based on findings from an optimization problem. Compared to an ad-hoc design of a heuristic, this has shown to be very time-consuming. It also increased the complexity of the work, because different methodologies were applied. This approach did, however, come with multiple advantages.

First, it allowed to identify and restrict the relevant variables. Thus, the final heuristic only has to adapt a single aspect of the system. Second, the preliminary evaluations facilitated the understanding of the overall system behavior. This allowed to avoid misjudgments. Instead of relying on estimations and guesses, the design of the heuristic thus became a straight-forward path. Third, this approach gave confidence that adapting the selected variables should result in the expected performance. In situations where preliminary implementations did not deliver this performance, this confidence allowed to concentrate on development, fixing, and tuning. In contrast, without the in-depth understanding of the system, a researcher could be tempted to hop between completely different, misleading approaches. Finally, the foundations laid with the initial optimization problem did here also serve for the final evaluation. As an assessment of absolute performance is often difficult in such complex systems, having a sound reference for relative comparison is beneficial.

## 7.3 Outlook

This thesis makes use of a model for the processing effort caused by signal processing for data transmission. The effort does, however highly depend on the chosen hardware architecture and software implementation of a C-RAN system. To apply the proposed mechanism, it is thus required to re-evaluate the results with a model correctly fitted to the target system. The model does furthermore not consider the cell specific processing (e. g. for OFDM and reference signals) and higher protocol layers (e. g. RLC and PDCP). While the higher layers can be assumed to have limited impact on the compute effort, the influence of the lower layers depends on the system architecture. In case they are performed at the BBU using the same processing resources as user specific calculations, it could be beneficial to include them in the evaluations. Such a system could, e. g., benefit from completely switching of a subset of transmit antennas for single cells.

This thesis focuses on processing effort for DL processing. However, UL processing also comes with high processing effort, and can be handled by the same processing units. To manage the computational effort for UL processing, Rost et al. [Ros+15b] proposed to use more robust encoding and thereby cope with fewer iterations in the turbo decoder. Similar to this thesis, they also selectively switch UEs to different MCSs. They do, however, propose a central algorithm for this task.

It remains open for further investigation whether the system proposed in this thesis could be combined with their approach. Such a system could, e. g., use a global variable to indicate the severity of the current overload. For DL operation, this is equivalent to the proposed threshold $e_{\min}$. For UL, the mapping of that variable to an MCS is performed independently for each UE. Finally, a fallback mechanism skips decoding for individual users in case the scheduled set of MCSs causes overload. The processing effort for UL operation not only depends on a correct compute effort model, but also on an accurate prediction of the channel quality. Thus, compared to DL, the prediction of effort for UL operation is more difficult. By jointly considering UL and DL processing, such a system can balance resources between both components.

Tight collaboration of adjacent cell sites, known as CoMP, is not covered by this thesis. CoMP comes with additional processing effort, because more transmit or receive antennas are combined for a single transmission. Thereby, the dimensioning of compute resources for the theoretical peak load becomes even more inefficient. It is possible to extend the proposed system to also include CoMP. For example, the decision to use CoMP to serve a UE can take a global processing load indicator into account. Thereby, in case processing resources become scarce, CoMP is only applied to those transmissions where it brings the most significant benefits. However, CoMP also influences RA in multiple cells, so it requires a tighter integration into the system.

This thesis assumes that the processing for multiple cells is performed by a single BBU pool. The resources of this pool are modeled as a single homogeneous mass. Such a setup is difficult to realize, because the performance of single processing units is limited. Larger computer systems achieve high performance by employing multiple compute units in parallel. Depending on the degree of coupling between these units, this has different impact on the proposed system.

In case tasks can be moved between compute units during a subframe, such an architecture can be covered by the proposed system. The encoding of data for UEs is performed in parallel. When the time reserved for processing is over, possibly multiple processing tasks have to be aborted. The impact on the expected performance is small as long as the number of compute units is significantly lower than the number of served UEs. When tasks cannot be moved between compute units instantly, an active load balancing is required, e. g. as proposed by Scholz and Grob-Lipski [SG16]. The compute units then act like separate, smaller BBU pools. Inside each pool, the variance of the requested compute load is higher, so more load peaks have to be cut off. This variance also makes the prediction of the global threshold more difficult. In reality, the architecture of a BBU pool probably lies between both concepts. Tighter coupling can be realized by employing larger computers and specialized interconnects.[1] As this often comes at a higher cost, an efficient design has to be found by jointly considering network performance and economical aspects.

[1] Inside a single computer, multiple cores use shared memory and can interact tightly. 20 to 40 cores are typical for current two socket x86 servers, e. g. based on Intel Broadwell architecture. 192 cores can be realized with an eight socket server, where each socket is equipped with an Intel Xeon E7-8894 v4 with 24 cores. Multiple of these computers can be interconnected with standard networking technology or dedicated low latency interconnects.

# A Definition of the Wrap-Around Geometry

The definition of the wrap-around geometry is based on the shape and rotation of the scenario. This geometry is designed to achieve a regularly repeating pattern of BS locations (see section 5.3.2). Two configurations of the geometry are defined. These correspond to two BS layouts with seven and 19 sites.

Assume that the distance between two adjacent BS sites (the *inter-site distance*) is defined as $d_{\text{is}}$. The shape of the scenario is a hexagon with an inscribing circle with radius $r_7 \approx 1.32 d_{\text{is}}$ or $r_{19} \approx 1.90 d_{\text{is}}$ for seven and 19 sites, respectively. For seven sites, the first edge points to $\alpha_7 \approx 49.11°$, for 19 sites to $\alpha_{19} \approx 53.41°$. For calculation of scenario sizes and rotations refer to equations (A.1) to (A.4). The whole scenario, including positions of the BSs, is depicted in figures A.1 and A.2.

$$r_7 = \sqrt{d_{\text{is}}^2 + \left(d_{\text{is}} \cos \frac{\pi}{6}\right)^2} = \frac{\sqrt{7}}{2} d_{\text{is}} \approx 1.32 d_{\text{is}} \tag{A.1}$$

$$r_{19} = \sqrt{(1.75 d_{\text{is}})^2 + \left(1.5 d_{\text{is}} \cos \frac{\pi}{6}\right)^2} = \frac{\sqrt{19}}{2} d_{\text{is}} \approx 2.18 d_{\text{is}} \tag{A.2}$$

$$\alpha_7 = \tan^{-1}\left(\frac{d_{\text{is}}}{d_{\text{is}} \cos \frac{\pi}{6}}\right) = \tan^{-1}\left(\frac{2}{\sqrt{3}}\right) \approx 49.11° \tag{A.3}$$

$$\alpha_{19} = \tan^{-1}\left(\frac{1.75 d_{\text{is}}}{1.5 d_{\text{is}} \cos \frac{\pi}{6}}\right) = \tan^{-1}\left(\frac{7}{3\sqrt{3}}\right) \approx 53.41° \tag{A.4}$$

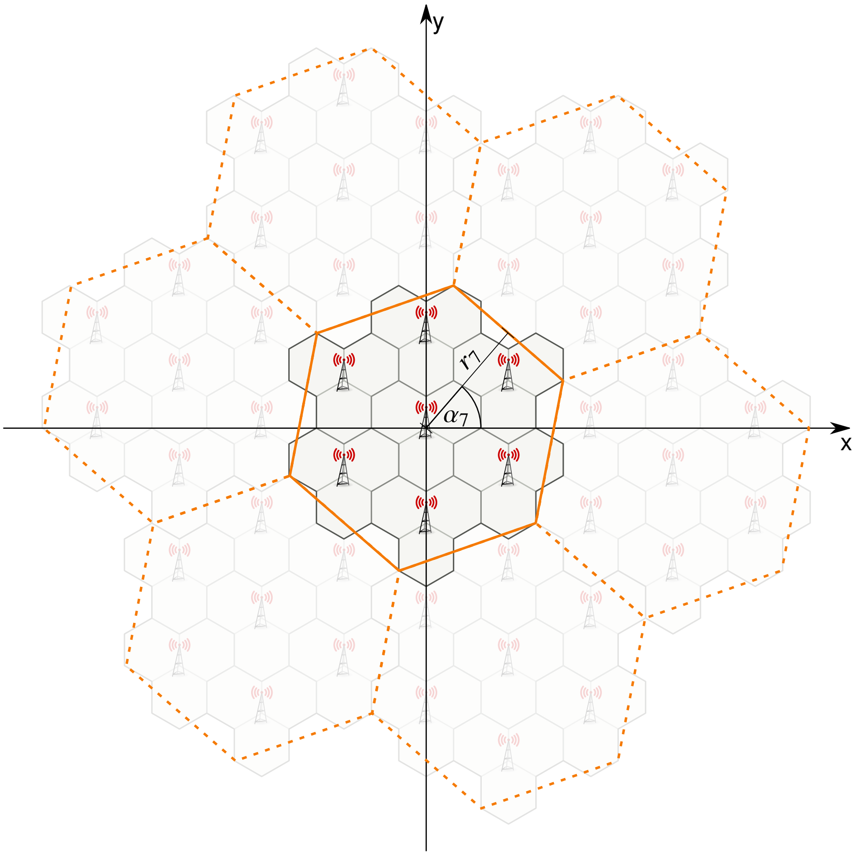


**Figure A.1:** Cell layout for seven tri-sectorized sites with wrap-around

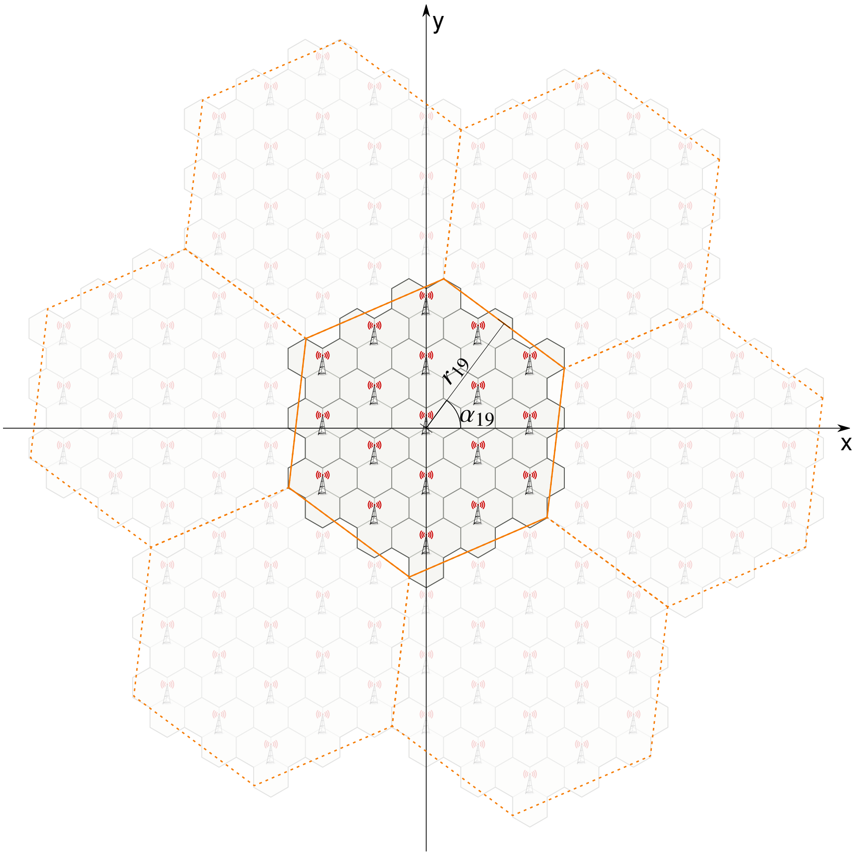


**Figure A.2:** Cell layout for 19 tri-sectorized sites with wrap-around

# B Measurement of Processing Effort

The objective of this appendix is to compare the processing effort model used in this thesis with a measurement available from literature.

Kai et al. [Kai+12] provide measurement results for a software implementation of a BBU. Their system uses one transmit antenna ($a = 1$, $l = 1$), 64-QAM modulation ($m = 6$), and code rate $^5\!/_6$ ($c = {}^5\!/_6$). It simultaneously transmits on 100 PRBs. Applying these parametrization to our model in equation (5.14) results in

$$\begin{aligned} P_{\mathrm{PRB}} &= 10^5 \left(3 \cdot 1 + 1 \cdot 1 + \frac{1}{3} \cdot 6 \cdot \frac{5}{6} \cdot 1\right) \\ &= 10^5 \left(3 + 1 + \frac{5}{3}\right) \\ &= 10^5 \left(\frac{17}{3}\right) \\ &= 567 \times 10^3. \end{aligned} \tag{B.1}$$

This is equivalent to a total effort of $100 \cdot P_{\mathrm{PRB}} = 56.7 \times 10^6$ operations per subframe.

The authors have measured a latency of 172 µs + 426 µs = 598 µs for symbol level and bit level operations, which corresponds to the processing tasks represented by our model. A single core of the Core i7-2600K processor with Sandy Bridge architecture performs eight *double precision* (DP) FLOPs per cycle (four multiplications, four additions) [Intel12]. With 3.4 GHz and four cores, this results in a total performance of $108.0 \times 10^9$ DP FLOPs per second, or $65 \times 10^6$ operations in 598 µs. This is about 115 % of the effort calculated with the model in equation (5.14).

Note that from the publication, it is not clear whether DP or *single precision* (SP) was used for the implementation. With SP, the same hardware can handle twice the number of operations per second. That results in approximately 230 % of the effort calculated by our model. However, the efficiency of signal processing greatly depends on the applied optimization.[1] Thereby, it can be assumed that the measured effort could be reduced by further optimization.

[1] For example, Tan et al. [Tan+11] realized speedups up to factor 50.

# C Tuning of the Proposed System

The proposed system has two tuning parameters $e_{\text{step}}$ and $p_{\text{off}}$, which were defined in section 5.6.3.4. These influence the control loop used to predict the efficiency threshold $e_{\text{min}}$. In this appendix, the configuration of these parameters is studied. Thereto, the same scenario is used as in section 6.1, i. e. a full-buffer traffic model with frequency selective RA and 100 ms simulated time.

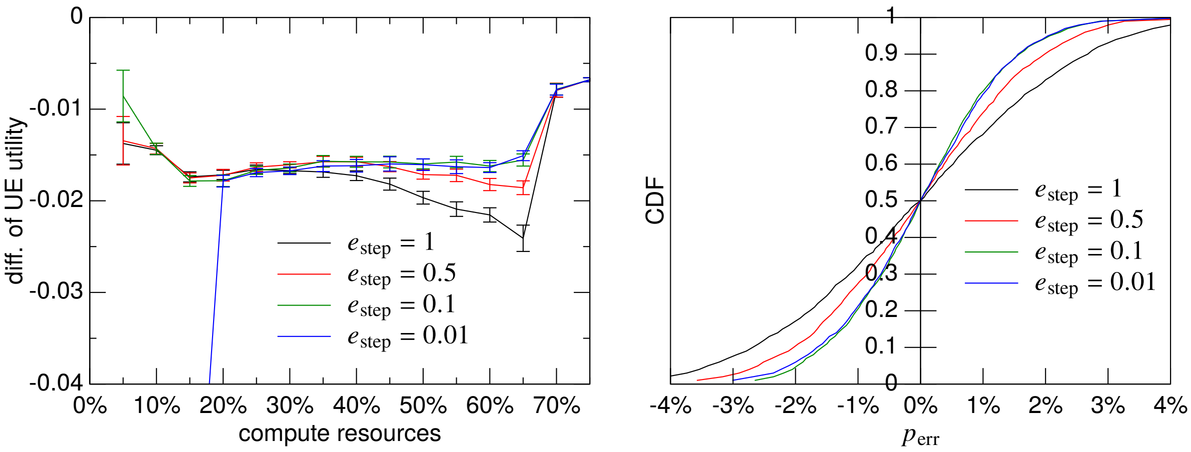
The main metric applied for these evaluations is the PF utility, i. e. the average of the logarithmized UE rates. This is used to avoid to be influenced by changes in the fairness, i. e. resources being shifted to cell border users or away from them. In contrast to the main evaluations in chapter 6, the focus of this appendix is not to provide easily interpretable metrics, but to facilitate a technical comparison. To be independent of the absolute value of the utility for a certain amount of compute resources, the utility is compared to the utility achieved with *opt. per subframe*, which was defined in section 6.2.2.3. This means that, for each amount of compute resources, the utility achieved with *opt. per subframe* is subtracted from all other utility values.[1]

As a second metric, the error of the total compute load resulting from the prediction is evaluated. This is calculated as $p_{\text{err}} = p_{\text{total}}^{\text{real}} - p_{\text{max}}$. In case $p_{\text{err}} > 0$, the system is in overload and has to skip the encoding for some UEs. Otherwise, i. e. whenever $p_{\text{err}} < 0$, compute resources are not fully utilized. To gain insight into the performance of the prediction component, the empirical CDF of this metric is plotted.

For the first study, the step size $e_{\text{step}}$ is evaluated while the offset is fixed at $p_{\text{off}} = 0$. Figures C.1a and C.1b show the difference of the UE utility and the error of the compute load, respectively. In figure C.1a can be seen that the step size has only marginal impact on the performance. In general, a large step size hinders the system to adapt closely to the optimal threshold. This is visible especially for 50 % to 60 % compute resources, where the optimal efficiency threshold is low. The lower the step size, the better the system can adapt to the optimal threshold. However, for very low thresholds and few available compute resources, the system does not reach the optimal threshold during the simulated warm-up time of 10 s.

The CDF of the error of the compute load supports this interpretation. Figure C.1b shows that all configurations have a median error of 0 %, i. e. they over- and underestimate the compute load with the same probability. For large step sizes, the curves become more flat, because there a close adaptation is not possible. The curves are steeper for smaller step sizes, but do not improve further for values of $e_{\text{step}} < 0.1$. This can be explained by the inherent variability of the

[1] The method used to calculate confidence intervals is the same as the one described in section 6.1.1 for relative differences. That means that, first, the difference in utility is calculated for each UE. These values are averaged per drop, and the confidence intervals are calculated from the drop-averages.



**(a)** difference of UE utility

**(b)** CDF of $p_{\text{err}}$ for $p_{\text{max}} = 60\,\%$

**Figure C.1:** Evaluation of the step size $e_{\text{step}}$ ($p_{\text{off}} = 0$)

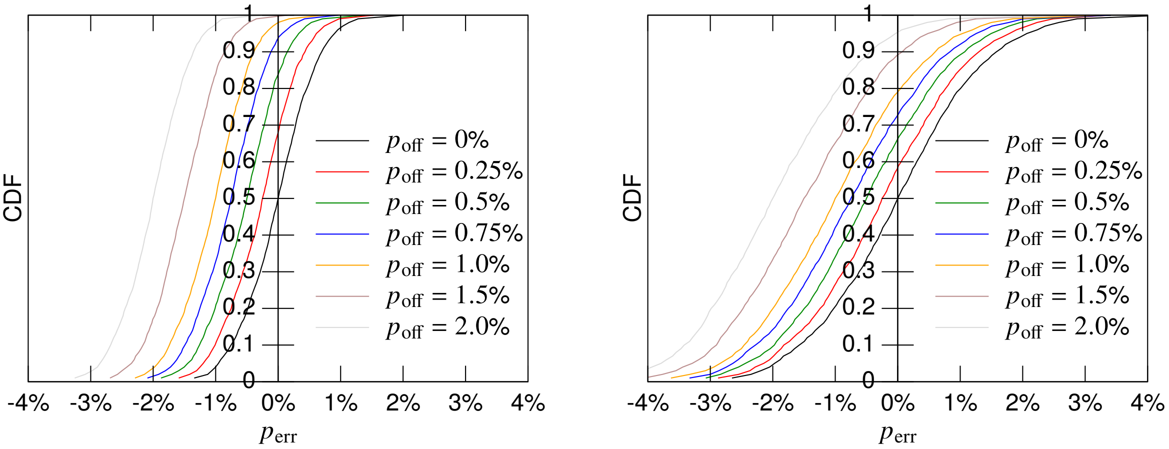
system, which is not covered by the prediction mechanism. Consequently, the threshold step size $e_{\text{step}} = 0.1$ is selected for the further studies.

The second study evaluates the effect of different values for the offset $p_{\text{off}}$, while the configuration of the step size is fixed to $e_{\text{step}} = 0.1$. Under constant conditions, $p_{\text{off}} = 0$ results in skipping a single UE every second subframe. A larger offset adds a safety margin, i. e. the resulting value of the threshold is higher. This results in a more restrictive MIMO mode selection, and consequently less compute resources are used. Thus, a larger offset can be used to avoid skipping at the cost of a reduced utilization of compute resources.

Figures C.2a and C.2b show the error of the compute load resulting from prediction for 20 % and 60 % compute resources, respectively. In general, the CDFs for $p_{\text{max}} = 20\,\%$ are steeper than those for $p_{\text{max}} = 60\,\%$. As expected, the CDFs of $p_{\text{err}}$ is shifted by the value of $-p_{\text{off}}$. Consequently, the occurrences of $p_{\text{err}} > 0$, i. e. overload situations which result in skipping, become more infrequent. With suitable configurations (e. g. $p_{\text{off}} = 1.5\,\%$ for $p_{\text{max}} = 20\,\%$), skipping can almost be avoided. At the same time, the available compute resources are utilized less efficiently. The area between the CDF, the $x$-axis (i. e., the lower border of the chart) and the $y$-axis (i. e., the vertical line at $p_{\text{err}} = 0$) can be interpreted as the wasted compute capacity.

The effect this shifting has on the PF utility can be seen in figure C.3. Whenever the system can use large amounts of compute resources, the error of the compute load is larger (the CDFs are less steep). Consequently, it benefits from moderate offset values. However, the same offset values result in reduced performance for $p_{\text{max}} < 40\,\%$. There, the error of the prediction is lower, and the effect of the inefficient utilization of the available resources dominates. When increasing the offset to values above 1.5 %, no improvement is visible for any compute resource limit.

Summarizing, there is no single optimal value of $p_{\text{off}}$. As a compromise, the value $p_{\text{off}} = 0.25\,\%$ is used for the studies in chapter 6. Compared to $p_{\text{off}} = 0\,\%$, this improves the utility for the range of 40 % to 70 % compute resources. At the same time, the drop of utility for less compute resources is not too severe. The design of a more elaborate control loop could facilitate a more efficient adaptation of the threshold over the whole range of compute resource limits.



**(a)** CDF of $p_{\text{err}}$ for $p_{\max} = 20\,\%$

**(b)** CDF of $p_{\text{err}}$ for $p_{\max} = 60\,\%$

**Figure C.2:** Evaluation of the offset $p_{\text{off}}$ ($e_{\text{step}} = 0.1$)

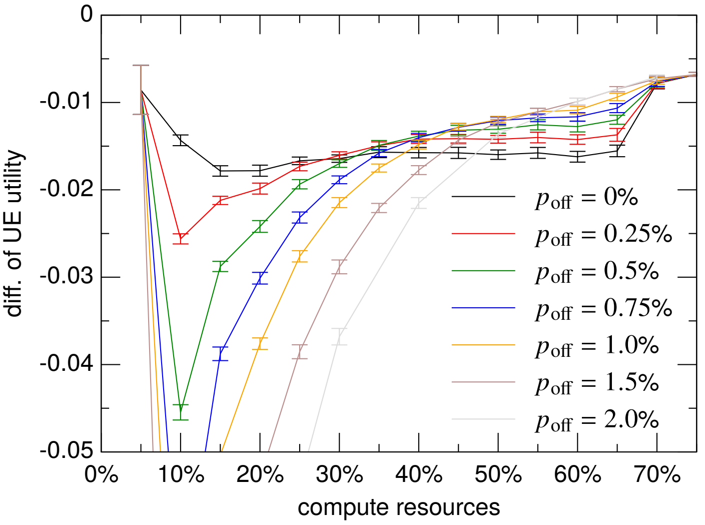


**Figure C.3:** Impact of the offset $p_{\text{off}}$ on the UE utility

# Bibliography

[3GPP 21.101 v8.4.0] John M. Meredith. *Technical Specifications and Technical Reports for a UTRAN-based 3GPP system*. Technical specification (TS) 21.101. Version 8.4.0. 3GPP, Mar. 2012.

[3GPP 23.203] Balazs Bertenyi. *Policy and charging control architecture*. Technical specification (TS) 23.203. 3GPP.

[3GPP 23.203 v10.10.0] Balazs Bertenyi. *Policy and charging control architecture*. Technical specification (TS) 23.203. Version 10.10.0. 3GPP, Dec. 2014.

[3GPP 23.401] Chris Pudney. *General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access*. Technical specification (TS) 23.401. 3GPP.

[3GPP 24.301] Jennifer Liu. *Non-Access-Stratum (NAS) protocol for Evolved Packet System (EPS); Stage 3*. Technical specification (TS) 24.301. 3GPP.

[3GPP 25.814] Sadayuki Abeta. *Physical layer aspect for evolved Universal Terrestrial Radio Access (UTRA)*. Technical report (TR) 25.814. 3GPP.

[3GPP 25.913] Takehiro Nakamura. *Requirements for Evolved UTRA (E-UTRA) and Evolved UTRAN (E-UTRAN)*. Technical report (TR) 25.913. 3GPP.

[3GPP 25.996] Howard Huang. *Spatial channel model for Multiple Input Multiple Output (MIMO) simulations*. Technical report (TR) 25.996. 3GPP.

[3GPP 36.104] Johan Skold. *Evolved Universal Terrestrial Radio Access (E-UTRA); Base Station (BS) radio transmission and reception*. Technical specification (TS) 36.104. 3GPP.

[3GPP 36.201] Matthew Baker. *Evolved Universal Terrestrial Radio Access (E-UTRA); LTE physical layer; General description*. Technical specification (TS) 36.201. 3GPP.

[3GPP 36.211] Stefan Parkvall. *Evolved Universal Terrestrial Radio Access (E-UTRA); Physical channels and modulation*. Technical specification (TS) 36.211. 3GPP.

[3GPP 36.212] Brian Classon. *Evolved Universal Terrestrial Radio Access (E-UTRA); Multiplexing and channel coding*. Technical specification (TS) 36.212. 3GPP.

[3GPP 36.213] Robert Love. *Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures*. Technical specification (TS) 36.213. 3GPP.

[3GPP 36.213 v12.5.0] Robert Love. *Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures*. Technical specification (TS) 36.213. Version 12.5.0. 3GPP, Mar. 2015.

[3GPP 36.300] Benoist Sebire. *Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2*. Technical specification (TS) 36.300. 3GPP.

[3GPP 36.300 v11.4.0] Benoist Sebire. *Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2*. Technical specification (TS) 36.300. Version 11.4.0. 3GPP, Dec. 2012.

[3GPP 36.300 v12.8.0] Benoist Sebire. *Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2*. Technical specification (TS) 36.300. Version 12.8.0. 3GPP, Dec. 2015.

[3GPP 36.300 v9.1.0] Benoist Sebire. *Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2*. Technical specification (TS) 36.300. Version 9.1.0. 3GPP, Sept. 2009.

[3GPP 36.321] Mats Folke. *Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) protocol specification*. Technical specification (TS) 36.321. 3GPP.

[3GPP 36.322] Toru Uchino. *Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Link Control (RLC) protocol specification*. Technical specification (TS) 36.322. 3GPP.

[3GPP 36.323] Seungjune Yi. *Evolved Universal Terrestrial Radio Access (E-UTRA); Packet Data Convergence Protocol (PDCP) specification*. Technical specification (TS) 36.323. 3GPP.

[3GPP 36.331] Himke van der Velde. *Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification*. Technical specification (TS) 36.331. 3GPP.

[3GPP 36.423] Markus Drevö. *Evolved Universal Terrestrial Radio Access Network (E-UTRAN); X2 Application Protocol (X2AP)*. Technical specification (TS) 36.423. 3GPP.

[3GPP 36.423 v10.1.0] Markus Drevö. *Evolved Universal Terrestrial Radio Access Network (E-UTRAN); X2 Application Protocol (X2AP)*. Technical specification (TS) 36.423. Version 10.1.0. 3GPP, Mar. 2011.

[3GPP 36.814] Sadayuki Abeta. *Evolved Universal Terrestrial Radio Access (E-UTRA); Further advancements for E-UTRA physical layer aspects*. Technical report (TR) 36.814. 3GPP.

[3GPP 36.819 v11.2.0] Bruno Clerckx. *Coordinated multi-point operation for LTE physical layer aspects*. Technical report (TR) 36.819. Version 11.2.0. 3GPP, Sept. 2013.

[3GPP 36.913] Takehiro Nakamura. *Requirements for further advancements for Evolved Universal Terrestrial Radio Access (E-UTRA) (LTE-A)*. Technical report (TR) 36.913. 3GPP.

[3GPP2 04] *cdma2000 Evaluation Methodology*. Tech. rep. C.R1002-0. Revision 0. 3GPP2, Dec. 2004.

[3GPP2 09] *cdma2000 Evaluation Methodology*. Tech. rep. C.R1002-A. Revision A. 3GPP2, May 2009.

[3GPPReleases] *Overview of 3GPP Releases*. URL: `http://www.3gpp.org/ftp/Information/WORK_PLAN/Description_Releases/` (visited on 01/08/2016).

[4GAm13] 4G Americas. *3GPP Release 11: Understanding the Standards for HSPA+ and LTE-Advanced Enhancements*. Aug. 2013.

[4GAm15] 4G Americas. *Understanding 3GPP Release 12: Standards for HSPA+ and LTE-Advanced Enhancements*. Feb. 2015.

[4GAm15b] 4G Americas. *Inside 3GPP Release 13: Understanding the Standards for HSPA+ and LTE-Advanced Enhancements*. Sept. 2015.

[5GAm17] 5G Americas. *Wireless Technology Evolution Towards 5G: 3GPP Release 13 to Release 15 and Beyond*. Feb. 2017.

[5GPPP16] *View on 5G Architecture*. Tech. rep. 5G PPP Architecture Working Group, July 2016.

[AAR12] Eitan Altman, Konstantin Avrachenkov, and Sreenath Ramanath. "Multiscale fairness and its application to resource allocation in wireless networks". In: *Computer Communications* 35.7 (2012), pp. 820–828. DOI: `http://dx.doi.org/10.1016/j.comcom.2012.01.013`.

[Ada97] A. Adas. "Traffic models in broadband networks". In: *IEEE Communications Magazine* 35.7 (July 1997), pp. 82–89. ISSN: 0163-6804. DOI: `10.1109/35.601746`.

[Ala98] S. M. Alamouti. "A simple transmit diversity technique for wireless communications". In: *IEEE Journal on Selected Areas in Communications* 16.8 (Oct. 1998), pp. 1451–1458. ISSN: 0733-8716. DOI: `10.1109/49.730453`.

[AM13] A. Asadi and V. Mancuso. "A Survey on Opportunistic Scheduling in Wireless Communications". In: *IEEE Communications Surveys Tutorials* 15.4 (2013), pp. 1671–1688. ISSN: 1553-877X. DOI: `10.1109/SURV.2013.011413.00082`.

[And+01] M. Andrews et al. "Providing quality of service over a shared wireless link". In: *IEEE Communications Magazine* 39.2 (Feb. 2001), pp. 150–154. ISSN: 0163-6804. DOI: `10.1109/35.900644`.

[And+14] J.G. Andrews et al. "What Will 5G Be?" In: *Selected Areas in Communications, IEEE Journal on* 32.6 (June 2014), pp. 1065–1082. ISSN: 0733-8716. DOI: `10.1109/JSAC.2014.2328098`.

[And05] J. G. Andrews. "Interference cancellation for cellular systems: a contemporary overview". In: *IEEE Wireless Communications* 12.2 (Apr. 2005), pp. 19–29. ISSN: 1536-1284. DOI: 10.1109/MWC.2005.1421925.

[And07] Matthew Andrews. "A Survey of Scheduling Theory in Wireless Data Networks". In: *Wireless Communications*. Ed. by Prathima Agrawal et al. New York, NY: Springer New York, 2007, pp. 1–17. ISBN: 978-0-387-48945-2. DOI: 10.1007/978-0-387-48945-2_1.

[APH13] O. El Ayach, S. W. Peters, and R. W. Heath. "The practical challenges of interference alignment". In: *IEEE Wireless Communications* 20.1 (Feb. 2013), pp. 35–42. ISSN: 1536-1284. DOI: 10.1109/MWC.2013.6472197.

[Årz+00] Karl-Erik Årzén et al. "An Introduction to Control and Scheduling Co-Design". eng. In: *Proceedings of the 39th IEEE Conference on Decision and Control, 2000.* Vol. 5. IEEE–Institute of Electrical and Electronics Engineers Inc., 2000, pp. 4865–4870. ISBN: 0-7803-6638-7. DOI: 10.1109/CDC.2001.914701.

[Ass08] M. Assaad. "Optimal Fractional Frequency Reuse (FFR) in Multicellular OFDMA System". In: *Vehicular Technology Conference, 2008. VTC 2008-Fall. IEEE 68th.* Sept. 2008, pp. 1–5. DOI: 10.1109/VETECF.2008.381.

[Bha+12] Sourjya Bhaumik et al. "CloudIQ: A Framework for Processing Base Stations in a Data Center". In: *Proceedings of the 18th Annual International Conference on Mobile Computing and Networking*. Mobicom '12. Istanbul, Turkey: ACM, 2012, pp. 125–136. ISBN: 978-1-4503-1159-5. DOI: 10.1145/2348543.2348561.

[BNetzA16] *Übersicht Mobilfunkspektrum nach der Auktion – Zuordnung ab 01.01.2017 gültig*. Tech. rep. Bundesnetzagentur, Jan. 2016. URL: http://www.bundesnetzagentur.de/SharedDocs/Downloads/DE/Sachgebiete/Telekommunikation/Unternehmen_Institutionen/Frequenzen/OffentlicheNetze/Mobilfunk/DrahtloserNetzzugang/Projekt2016/Frequenzen700bis1800_pdf.pdf (visited on 04/22/2016).

[Bru+05] K. Brueninghaus et al. "Link performance models for system level simulations of broadband radio access systems". In: *2005 IEEE 16th International Symposium on Personal, Indoor and Mobile Radio Communications*. Vol. 4. Sept. 2005, pp. 2306–2311. DOI: 10.1109/PIMRC.2005.1651855.

[But+10] Giorgio Buttazzo et al. *Soft Real-Time Systems*. Springer, 2010. ISBN: 978-1-4419-3655-4.

[Cap+13] F. Capozzi et al. “Downlink Packet Scheduling in LTE Cellular Networks: Key Design Issues and a Survey”. In: *IEEE Communications Surveys Tutorials* 15.2 (June 2013), pp. 678–700. ISSN: 1553-877X. DOI: 10.1109/SURV.2012.060912.00100.

[Car+13] F.D. Cardoso et al. “Energy efficient transmission techniques for LTE”. In: *Communications Magazine, IEEE* 51.10 (Oct. 2013), pp. 182–190. ISSN: 0163-6804. DOI: 10.1109/MCOM.2013.6619582.

[CGB04] Shuguang Cui, A. J. Goldsmith, and A. Bahai. “Energy-efficiency of MIMO and cooperative MIMO techniques in sensor networks”. In: *IEEE Journal on Selected Areas in Communications* 22.6 (Aug. 2004), pp. 1089–1098. ISSN: 0733-8716. DOI: 10.1109/JSAC.2004.830916.

[CGB05] Shuguang Cui, A. J. Goldsmith, and A. Bahai. “Energy-constrained modulation optimization”. In: *IEEE Transactions on Wireless Communications* 4.5 (Sept. 2005), pp. 2349–2360. ISSN: 1536-1276. DOI: 10.1109/TWC.2005.853882.

[Cha+02] Etienne F. Chaponniere et al. “Transmitter Directed Code Division Multiple Access System Using Path Diversity To Equitably Maximize Throughput”. Pat. 6449490 B1 (US). Sept. 2002.

[Che+14] A. Checko et al. “Cloud RAN for Mobile Networks - a Technology Overview”. In: *Communications Surveys Tutorials, IEEE* PP.99 (2014), pp. 1–1. ISSN: 1553-877X. DOI: 10.1109/COMST.2014.2355255.

[CJ08] V. R. Cadambe and S. A. Jafar. “Interference Alignment and Degrees of Freedom of the K -User Interference Channel”. In: *IEEE Transactions on Information Theory* 54.8 (Aug. 2008), pp. 3425–3441. ISSN: 0018-9448. DOI: 10.1109/TIT.2008.926344.

[CL01] Yaxin Cao and V. O. K. Li. “Scheduling algorithms in broadband wireless networks”. In: *Proceedings of the IEEE* 89.1 (Jan. 2001), pp. 76–87. ISSN: 0018-9219. DOI: 10.1109/5.904507.

[CMRI11] *C-RAN The Road Towards Green RAN*. Tech. rep. v2.5. China Mobile Research Institute, 2011. URL: http://labs.chinamobile.com/cran/wp-content/uploads/CRAN_white_paper_v2_5_EN.pdf (visited on 08/04/2016).

[Col13] J. Colom Ikuno. “System Level Modeling and Optimization of the LTE Downlink”. PhD thesis. Vienna: Vienna University of Technology, 2013.

[Cor+10] L. M. Correia et al. “Challenges and enabling technologies for energy aware mobile radio networks”. In: *IEEE Communications Magazine* 48.11 (Nov. 2010), pp. 66–72. ISSN: 0163-6804. DOI: 10.1109/MCOM.2010.5621969.

[Cos83] M. Costa. “Writing on dirty paper (Corresp.)” In: *IEEE Transactions on Information Theory* 29.3 (May 1983), pp. 439–441. ISSN: 0018-9448. DOI: 10.1109/TIT.1983.1056659.

[Cox14] Christopher I. Cox. *An Introduction to LTE: LTE, LTE-Advanced, SAE, VoLTE and 4G Mobile Communications*. 2nd ed. Chichester: Wiley, July 2014. ISBN: 9781118818015.

[CS03] G. Caire and S. Shamai. “On the achievable throughput of a multiantenna Gaussian broadcast channel”. In: *IEEE Transactions on Information Theory* 49.7 (July 2003), pp. 1691–1706. ISSN: 0018-9448. DOI: 10.1109/TIT.2003.813523.

[CS55] D. R. Cox and A. Stuart. “Some Quick Sign Tests for Trend in Location and Dispersion”. In: *Biometrika* 42.1/2 (1955), pp. 80–95.

[CTL12] Tzi-Dar Chiueh, Pei-Yun Tsai, and I-Wei Lai. *Baseband receiver design for wireless MIMO-OFDM communications*. John Wiley & Sons, 2012.

[DDL13] C. Desset, B. Debaillie, and F. Louagie. “Towards a flexible and future-proof power model for cellular base stations”. In: *Digital Communications - Green ICT (TIWDC), 2013 24th Tyrrhenian International Workshop on*. Sept. 2013, pp. 1–6. DOI: 10.1109/TIWDC.2013.6664200.

[DDL14] C. Desset, B. Debaillie, and F. Louagie. “Modeling the hardware power consumption of large scale antenna systems”. In: *Green Communications (OnlineGreencomm), 2014 IEEE Online Conference on*. Nov. 2014, pp. 1–6. DOI: 10.1109/OnlineGreenCom.2014.7114430.

[Deb+14] Supratim Deb et al. “Algorithms for Enhanced Inter-Cell Interference Coordination (eICIC) in LTE HetNets”. In: *IEEE/ACM Trans. Netw.* 22.1 (Feb. 2014), pp. 137–150. ISSN: 1063-6692. DOI: 10.1109/TNET.2013.2246820.

[Des+11] Claude Desset et al. “Chapter 6 - Implementing Scalable List Detectors for MIMO-SDM in LTE”. In: *MIMO*. Ed. by Alain Sibille, Claude Oestges, and Alberto Zanella. Oxford: Academic Press, 2011, pp. 175–193. ISBN: 978-0-12-382194-2. DOI: 10.1016/B978-0-12-382194-2.00006-X.

[Des+12] C. Desset et al. “Flexible power modeling of LTE base stations”. In: *Wireless Communications and Networking Conference (WCNC), 2012 IEEE*. Apr. 2012, pp. 2858–2862. DOI: 10.1109/WCNC.2012.6214289.

[DKS89] A. Demers, S. Keshav, and S. Shenker. “Analysis and Simulation of a Fair Queueing Algorithm”. In: *SIGCOMM Comput. Commun. Rev.* 19.4 (Aug. 1989), pp. 1–12. ISSN: 0146-4833. DOI: 10.1145/75247.75248.

[DPS14] E. Dahlman, S. Parkvall, and J. Sköld. *4G LTE/LTE-Advanced for Mobile Broadband*. second. Oxford: Elsevier, 2014.

[DPS16] E. Dahlman, S. Parkvall, and J. Sköld. *4G LTE-Advanced Pro and The Road to 5G*. third. Oxford: Elsevier, 2016.

[DT12] Claude Desset and Rodolfo Torrea Duran. "Reducing the power of wireless terminals by adaptive baseband processing". In: *Annals of Telecommunications - Annales des Télécommunications* 67.3-4 (Apr. 2012), pp. 161–170.

[DVR03] Suman Das, Harish Viswanathan, and G. Rittenhouse. "Dynamic load balancing through coordinated scheduling in packet data systems". In: *INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications. IEEE Societies*. Vol. 1. Mar. 2003, 786–796 vol.1. DOI: 10.1109/INFCOM.2003.1208728.

[EARTH] *EARTH (Energy Aware Radio and neTwork tecHnologies)*. EU Funded Research Project, Jan. 2010 – June 2012. URL: https://www.ict-earth.eu (visited on 08/03/2016).

[Ell+09] J. Ellenbeck et al. "A Concept for Efficient System-Level Simulations of OFDMA Systems with Proportional Fair Fast Scheduling". In: *2009 IEEE Globecom Workshops*. Nov. 2009, pp. 1–6. DOI: 10.1109/GLOCOMW.2009.5360729.

[FL02] H. Fattah and C. Leung. "An overview of scheduling algorithms in wireless multimedia networks". In: *IEEE Wireless Communications* 9.5 (Oct. 2002), pp. 76–83. ISSN: 1536-1284. DOI: 10.1109/MWC.2002.1043857.

[Flo15] Dino Flore. *RAN workshop on 5G: Chairman Summary*. Sept. 2015.

[GCL14] Marisol García-Valls, Tommaso Cucinotta, and Chenyang Lu. "Challenges in real-time virtualization and predictable cloud computing". In: *Journal of Systems Architecture* 60.9 (2014), pp. 726–740. ISSN: 1383-7621. DOI: 10.1016/j.sysarc.2014.07.004.

[Ges+07] D. Gesbert et al. "Adaptation, Coordination, and Distributed Resource Allocation in Interference-Limited Wireless Networks". In: *Proceedings of the IEEE* 95.12 (Dec. 2007), pp. 2393–2409. ISSN: 0018-9219. DOI: 10.1109/JPROC.2007.907125.

[GHL93] Alan Garvey, Marty Humphrey, and Victor Lesser. "Task Interdependencies in Design-to-time Real-time Scheduling". In: *Proceedings of the Eleventh National Conference on Artificial Intelligence*. AAAI'93. Washington, D.C.: AAAI Press, 1993, pp. 580–585. ISBN: 0-262-51071-5.

[Gje+06] A. Gjendemsjo et al. "Optimal Power Allocation and Scheduling for Two-Cell Capacity Maximization". In: *2006 4th International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks*. Apr. 2006, pp. 1–6. DOI: 10.1109/WIOPT.2006.1666517.

[GJS03] J. Gozdecki, A. Jajszczyk, and R. Stankiewicz. "Quality of service terminology in IP networks". In: *IEEE Communications Magazine* 41.3 (Mar. 2003), pp. 153–159. ISSN: 0163-6804. DOI: 10.1109/MCOM.2003.1186560.

[Gra+79] Ronald L Graham et al. "Optimization and approximation in deterministic sequencing and scheduling: a survey". In: *Annals of discrete mathematics* 5 (1979), pp. 287–326.

[Gua+16] Fei Guan et al. "Open source FreeRTOS as a case study in real-time operating system evolution". In: *Journal of Systems and Software* 118 (2016), pp. 19–35. ISSN: 0164-1212. DOI: 10.1016/j.jss.2016.04.063.

[HBB11] Z. Hasan, H. Boostanimehr, and V. K. Bhargava. "Green Cellular Networks: A Survey, Some Research Issues and Challenges". In: *IEEE Communications Surveys Tutorials* 13.4 (2011), pp. 524–540. ISSN: 1553-877X. DOI: 10.1109/SURV.2011.092311.00031.

[HBH06] Jianwei Huang, R. A. Berry, and M. L. Honig. "Distributed interference compensation for wireless networks". In: *IEEE Journal on Selected Areas in Communications* 24.5 (May 2006), pp. 1074–1084. ISSN: 0733-8716. DOI: 10.1109/JSAC.2006.872889.

[HL08] Zhu Han and K. J. Ray Liu. *Resource Allocation for Wireless Networks – Basics, Techniques, and Applications*. Cambridge, UK: Cambridge University Press, Sept. 2008.

[Hua+09] J. Huang et al. "Downlink scheduling and resource allocation for OFDM systems". In: *IEEE Transactions on Wireless Communications* 8.1 (Jan. 2009), pp. 288–296. ISSN: 1536-1276. DOI: 10.1109/T-WC.2009.071266.

[I+14a] Chih-Lin I et al. "Recent Progress on C-RAN Centralization and Cloudification". In: *Access, IEEE* 2 (2014), pp. 1030–1039. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2014.2351411.

[I+14b] Chih-Lin I et al. "Toward green and soft: a 5G perspective". In: *Communications Magazine, IEEE* 52.2 (Feb. 2014), pp. 66–73. ISSN: 0163-6804. DOI: 10.1109/MCOM.2014.6736745.

[IKRSimLib] Institute of Communication Networks and Computer Engineering (IKR), University of Stuttgart. *IKR Simulation and Emulation Library (IKR SimLib)*. online, visited on 10/11/2017. URL: https://www.ikr.uni-stuttgart.de/institute/infrastructure/ikr-simlib-object-oriented-simulation-library/.

[Intel12] *Intel 64 and IA-32 Architectures Optimization Reference Manual*. Apr. 2012.

[Jan+11] J. Janhunen et al. "Fixed- and Floating-Point Processor Comparison for MIMO-OFDM Detector". In: *Selected Topics in Signal Processing, IEEE Journal of* 5.8 (Dec. 2011), pp. 1588–1598. ISSN: 1932-4553. DOI: 10.1109/JSTSP.2011.2165830.

[Jay04] Sudharman K Jayaweera. "Energy analysis of MIMO techniques in wireless sensor networks". In: *38th conference on information sciences and systems*. 2004.

[JCH84] Raj Jain, Dah-Ming Chiu, and William R. Hawe. *A quantitative measure of fairness and discrimination for resource allocation in shared computer system*. Tech. rep. Hudson, MA: Eastern Research Laboratory, Digital Equipment Corporation, 1984.

[JGL05] Z. Jiang, Y. Ge, and Y. Li. “Max-utility wireless resource management for best-effort traffic”. In: *IEEE Transactions on Wireless Communications* 4.1 (Jan. 2005), pp. 100–111. ISSN: 1536-1276. DOI: 10.1109/TWC.2004.840210.

[JPP00] A. Jalali, R. Padovani, and R. Pankaj. “Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system”. In: *Vehicular Technology Conference Proceedings, 2000. VTC 2000-Spring Tokyo. 2000 IEEE 51st*. Vol. 3. 2000, 1854–1858 vol.3. DOI: 10.1109/VETECS.2000.851593.

[Kai+12] Niu Kai et al. “LTE eNodeB prototype based on GPP platform”. In: *Globecom Workshops, 2012 IEEE*. Dec. 2012, pp. 279–284. DOI: 10.1109/GLOCOMW.2012.6477583.

[Kas+10] M. Kaschub et al. “Interference mitigation by distributed beam forming optimization”. In: *Frequenz - Journal of RF-Engineering and Telecommunications, Special Issue: "Interference Management and Cooperation Strategies in Communication Networks"* (Sept. 2010).

[Kas16] Matthias J. Kaschub. “Quality of Transaction”. PhD thesis. Institute of Communication Networks and Computer Engineering, University of Stuttgart, May 2016.

[Kel+08] P. Kela et al. “Dynamic packet scheduling performance in UTRA Long Term Evolution downlink”. In: *Wireless Pervasive Computing, 2008. ISWPC 2008. 3rd International Symposium on*. May 2008, pp. 308–313. DOI: 10.1109/ISWPC.2008.4556220.

[Kel97] Frank Kelly. “Charging and rate control for elastic traffic”. In: *European transactions on Telecommunications* 8.1 (1997), pp. 33–37.

[KHK04] Hoon Kim, Youngnam Han, and Jayong Koo. “Optimal subchannel allocation scheme in multicell OFDMA systems”. In: *Vehicular Technology Conference, 2004. VTC 2004-Spring. 2004 IEEE 59th*. Vol. 3. May 2004, 1821–1825 Vol.3. DOI: 10.1109/VETECS.2004.1390571.

[Kim+09] H. Kim et al. “A cross-layer approach to energy efficiency for adaptive MIMO systems exploiting spare capacity”. In: *IEEE Transactions on Wireless Communications* 8.8 (Aug. 2009), pp. 4264–4275. ISSN: 1536-1276. DOI: 10.1109/TWC.2009.081123.

[KMN02] K. Keutzer, S. Malik, and A. R. Newton. “From ASIC to ASIP: the next design discontinuity”. In: *Computer Design: VLSI in Computers and Processors, 2002. Proceedings. 2002 IEEE International Conference on*. 2002, pp. 84–90. DOI: 10.1109/ICCD.2002.1106752.

[KPP04] Hans Kellerer, Ulrich Pferschy, and David Pisinger. *Knapsack Problems*. Berlin: Springer, 2004.

[KWM11] F. Kienle, N. Wehn, and H. Meyr. "On Complexity, Energy- and Implementation-Efficiency of Channel Decoders". In: *IEEE Transactions on Communications* 59.12 (Dec. 2011), pp. 3301–3310. ISSN: 0090-6778. DOI: 10.1109/TCOMM.2011.092011.100157.

[Law07] Averill M. Law. *Simulation Modeling and Analysis*. Fourth edition. McGraw-Hill, 2007.

[Lee+12] Daewon Lee et al. "Coordinated multipoint transmission and reception in LTE-advanced: deployment scenarios and operational challenges". In: *Communications Magazine, IEEE* 50.2 (Feb. 2012), pp. 148–155. ISSN: 0163-6804. DOI: 10.1109/MCOM.2012.6146494.

[Li+11] Min Li et al. "Energy Aware Signal Processing for Software Defined Radio Baseband Implementation". English. In: *Journal of Signal Processing Systems* 63.1 (2011), pp. 13–25. ISSN: 1939-8018. DOI: 10.1007/s11265-009-0359-y.

[Lin+10] Y. Lin et al. "Wireless network cloud: Architecture and system requirements". In: *IBM Journal of Research and Development* 54.1 (Jan. 2010), 4:1–4:12. ISSN: 0018-8646. DOI: 10.1147/JRD.2009.2037680.

[Liu+91] J. W. S. Liu et al. "Algorithms for scheduling imprecise computations". In: *Computer* 24.5 (May 1991), pp. 58–68. ISSN: 0018-9162. DOI: 10.1109/2.76287.

[Liu+94] J. W. S. Liu et al. "Imprecise computations". In: *Proceedings of the IEEE* 82.1 (Jan. 1994), pp. 83–94. ISSN: 0018-9219. DOI: 10.1109/5.259428.

[LL03] Guoqing Li and Hui Liu. "Downlink dynamic resource allocation for multi-cell OFDMA system". In: *Vehicular Technology Conference, 2003. VTC 2003-Fall. 2003 IEEE 58th.* Vol. 3. Oct. 2003, 1698–1702 Vol.3. DOI: 10.1109/VETECF.2003.1285314.

[Lop+11] D. Lopez-Perez et al. "Enhanced intercell interference coordination challenges in heterogeneous networks". In: *IEEE Wireless Communications* 18.3 (June 2011), pp. 22–30. ISSN: 1536-1284. DOI: 10.1109/MWC.2011.5876497.

[Lov+08] D. J. Love et al. "An overview of limited feedback in wireless communication systems". In: *IEEE Journal on Selected Areas in Communications* 26.8 (Oct. 2008), pp. 1341–1365. ISSN: 0733-8716. DOI: 10.1109/JSAC.2008.081002.

[LSS06] Xiaojun Lin, N. B. Shroff, and R. Srikant. "A tutorial on cross-layer optimization in wireless networks". In: *IEEE Journal on Selected Areas in Communications* 24.8 (Aug. 2006), pp. 1452–1463. ISSN: 0733-8716. DOI: 10.1109/JSAC.2006.879351.

[LV89] R. Lupas and S. Verdu. "Linear multiuser detectors for synchronous code-division multiple-access channels". In: *IEEE Transactions on Information Theory* 35.1 (Jan. 1989), pp. 123–136. ISSN: 0018-9448. DOI: 10.1109/18.42183.

[M.1457-12] *Detailed specifications of the terrestrial radio interfaces of International Mobile Telecommunications-2000 (IMT-2000)*. Recommendation. ITU-R, Feb. 2015.

[M.2012-0] *Detailed specifications of the terrestrial radio interfaces of International Mobile Telecommunications-Advanced (IMT-Advanced)*. Recommendation. ITU-R, Jan. 2012.

[M.2083-0] *IMT Vision – Framework and overall objectives of the future development of IMT for 2020 and beyond*. Recommendation. ITU-R, Sept. 2015.

[Maa+12] Helka-Liina Maattanen et al. "System-level performance of LTE-Adv. with joint transmission and dynamic point selection schemes". In: *EURASIP Journal on Advances in Signal Processing* 2012.1 (2012), p. 247. ISSN: 1687-6180. DOI: 10.1186/1687-6180-2012-247.

[Mad+10] R. Madan et al. "Cell Association and Interference Coordination in Heterogeneous LTE-A Cellular Networks". In: *IEEE Journal on Selected Areas in Communications* 28.9 (Dec. 2010), pp. 1479–1489. ISSN: 0733-8716. DOI: 10.1109/JSAC.2010.101209.

[Mar+16] Robert Margolies et al. "Exploiting Mobility in Proportional Fair Cellular Scheduling: Measurements and Algorithms". In: *IEEE/ACM Trans. Netw.* 24.1 (Feb. 2016), pp. 355–367. ISSN: 1063-6692. DOI: 10.1109/TNET.2014.2362928.

[Meh+11] Christian Mehlführer et al. "The Vienna LTE simulators - Enabling reproducibility in wireless communications research". In: *Advances in Signal Processing, EURASIP Journal on* (July 2011). DOI: 10.1186/1687-6180-2011-29.

[MF11] Patrick Marsch and Gerhard P. Fettweis. *Coordinated Multi-Point in Mobile Communications: From Theory to Practice*. Cambridge University Press, 2011.

[Mie+09] J. Mietzner et al. "Multiple-antenna techniques for wireless communications - a comprehensive literature survey". In: *Communications Surveys Tutorials, IEEE* 11.2 (2009), pp. 87–105. ISSN: 1553-877X. DOI: 10.1109/SURV.2009.090207.

[MMK08] M. A. Maddah-Ali, A. S. Motahari, and A. K. Khandani. "Communication Over MIMO X Channels: Interference Alignment, Decomposition, and Performance Analysis". In: *IEEE Transactions on Information Theory* 54.8 (Aug. 2008), pp. 3457–3470. ISSN: 0018-9448. DOI: 10.1109/TIT.2008.926460.

[Mon+08] G. Monghal et al. "QoS Oriented Time and Frequency Domain Packet Schedulers for The UTRAN Long Term Evolution". In: *Vehicular Technology Conference, 2008. VTC Spring 2008. IEEE*. May 2008, pp. 2532–2536. DOI: 10.1109/VETECS.2008.557.

[Myu07] H. G. Myung. "Introduction to single carrier FDMA". In: *Signal Processing Conference, 2007 15th European*. Sept. 2007, pp. 2144–2148.

[Nec09] M.C. Necker. “A Novel Algorithm for Distributed Dynamic Interference Coordination in Cellular OFDMA Networks - Communication Networks and Computer Engineering Report No. 101”. PhD thesis. Universität Stuttgart, 2009.

[NGM08] NGMN Alliance. *NGMN Radio Access Performance Evaluation Methodology*. Ed. by Ralf Irmer. Jan. 2008.

[Oli+16] A. de la Oliva et al. “An overview of the CPRI specification and its application to C-RAN-based LTE scenarios”. In: *Communications Magazine, IEEE* 54.2 (Feb. 2016), pp. 152–159. ISSN: 0163-6804. DOI: `10.1109/MCOM.2016.7402275`.

[PH94] P. Patel and J. Holtzman. “Analysis of a simple successive interference cancellation scheme in a DS/CDMA system”. In: *IEEE Journal on Selected Areas in Communications* 12.5 (June 1994), pp. 796–807. ISSN: 0733-8716. DOI: `10.1109/49.298053`.

[PHT16] D. Pompili, A. Hajisami, and T.X. Tran. “Elastic resource utilization framework for high capacity and energy efficiency in cloud RAN”. In: *Communications Magazine, IEEE* 54.1 (Jan. 2016), pp. 26–32. ISSN: 0163-6804. DOI: `10.1109/MCOM.2016.7378422`.

[Pic+08] M. Pickavet et al. “Worldwide energy needs for ICT: The rise of power-aware networking”. In: *2008 2nd International Symposium on Advanced Networks and Telecommunication Systems*. Dec. 2008, pp. 1–3. DOI: `10.1109/ANTS.2008.4937762`.

[PKV11] M. Proebster, M. Kaschub, and S. Valentin. “Context-Aware Resource Allocation to Improve the Quality of Service of Heterogeneous Traffic”. In: *IEEE International Conference on Communications (ICC 2011)*. June 2011.

[Pro+12] M. Proebster et al. “Context-Aware Resource Allocation for Cellular Wireless Networks”. In: *EURASIP Journal on Wireless Communications and Networking* (2012).

[Pro15] Magnus Christian Proebster. “Size-Based Scheduling to Improve the User Experience in Cellular Networks”. PhD thesis. Institute of Communication Networks and Computer Engineering, University of Stuttgart, June 2015.

[RHK03] Jong-Hun Rhee, J. M. Holtzman, and Dong-Ku Kim. “Scheduling of real/non-real time services: adaptive EXP/PF algorithm”. In: *Vehicular Technology Conference, 2003. VTC 2003-Spring. The 57th IEEE Semiannual*. Vol. 1. Apr. 2003, 462–466 vol.1. DOI: `10.1109/VETECS.2003.1207583`.

[Rin47] Douglas H. Ring. *Mobile Telephony – Wide Area Coverage*. Tech. rep. Dec. 1947.

[Ros+14] P. Rost et al. “Cloud technologies for flexible 5G radio access networks”. In: *Communications Magazine, IEEE* 52.5 (May 2014), pp. 68–76. ISSN: 0163-6804. DOI: `10.1109/MCOM.2014.6898939`.

[Ros+15a] P. Rost et al. “Benefits and challenges of virtualization in 5G radio access networks”. In: *Communications Magazine, IEEE* 53.12 (Dec. 2015), pp. 75–82. ISSN: 0163-6804. DOI: 10.1109/MCOM.2015.7355588.

[Ros+15b] P. Rost et al. “Computationally Aware Sum-Rate Optimal Scheduling for Centralized Radio Access Networks”. In: *2015 IEEE Global Communications Conference (GLOBECOM)*. Dec. 2015, pp. 1–6. DOI: 10.1109/GLOCOM.2015.7417498.

[RST16] Markus Rupp, Stefan Schwarz, and Martin Taranetz. *The Vienna LTE-Advanced Simulators: Up and Downlink, Link and System Level Simulation*. 1st ed. Signals and Communication Technology. Springer Singapore, 2016. ISBN: 978-981-10-0616-6. DOI: 10.1007/978-981-10-0617-3.

[RTV15] P. Rost, S. Talarico, and M. C. Valenti. “The Complexity-Rate Tradeoff of Centralized Radio Access Networks”. In: *IEEE Transactions on Wireless Communications* 14.11 (Nov. 2015), pp. 6164–6176. ISSN: 1536-1276. DOI: 10.1109/TWC.2015.2449321.

[RY10] M. Rahman and H. Yanikomeroglu. “Enhancing cell-edge performance: a downlink dynamic interference avoidance scheme with inter-cell coordination”. In: *IEEE Transactions on Wireless Communications* 9.4 (Apr. 2010), pp. 1414–1425. ISSN: 1536-1276. DOI: 10.1109/TWC.2010.04.090256.

[SAE05] Zukang Shen, Jeffrey G Andrews, and Brian L Evans. “Adaptive resource allocation in multiuser OFDM systems with proportional rate constraints”. In: *Wireless Communications, IEEE Transactions on* 4.6 (2005), pp. 2726–2737.

[Sal+05] Jari Salo et al. MATLAB *implementation of the 3GPP Spatial Channel Model [3GPP 25.996]*. archived by http://web.archive.org on 2010-04-05. Jan. 2005. URL: http://radio.aalto.fi/en/research/rf_applications_in_mobile_communications/propagation_research/matlab_scm_implementation/ (visited on 10/28/2014).

[SAR09] S. Sadr, A. Anpalagan, and K. Raahemifar. “Radio Resource Allocation Algorithms for the Downlink of Multiuser OFDM Communication Systems”. In: *IEEE Communications Surveys Tutorials* 11.3 (Aug. 2009), pp. 92–106. ISSN: 1553-877X. DOI: 10.1109/SURV.2009.090307.

[Sen+06] Gamini Senarath et al. *Multi-hop Relay System Evaluation Methodology (Channel Model and Performance Metric)*. IEEE 802.16 Broadband Wireless Access Working Group, Sept. 2006.

[SG16] S. Scholz and H. Grob-Lipski. “Reallocation Strategies for User Processing Tasks in Future Cloud-RAN Architectures”. In: *Proceedings of the 27th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC 2016)*. 2016.

[Sha+04] Lui Sha et al. “Real Time Scheduling Theory: A Historical Perspective”. In: *Real-Time Syst.* 28.2-3 (Nov. 2004), pp. 101–155. ISSN: 0922-6443. DOI: 10.1023/B:TIME.0000045315.61234.1e.

[Sha49] Claude E Shannon. “Communication in the presence of noise”. In: *Proceedings of the IRE* 37.1 (1949), pp. 10–21.

[She95] S. Shenker. “Fundamental design issues for the future Internet”. In: *IEEE Journal on Selected Areas in Communications* 13.7 (Sept. 1995), pp. 1176–1188. ISSN: 0733-8716. DOI: 10.1109/49.414637.

[Shi+89] W. Shih et al. “Scheduling Tasks with Ready Times and Deadlines to Minimize Average Error”. In: *SIGOPS Oper. Syst. Rev.* 23.3 (July 1989), pp. 14–28. ISSN: 0163-5980. DOI: 10.1145/71021.71022.

[SJT09] C. So-In, R. Jain, and A. K. Tamimi. “Scheduling in IEEE 802.16e mobile WiMAX networks: key issues and a survey”. In: *IEEE Journal on Selected Areas in Communications* 27.2 (Feb. 2009), pp. 156–171. ISSN: 0733-8716. DOI: 10.1109/JSAC.2009.090207.

[Skl+16] G. Sklivanitis et al. “Addressing next-generation wireless challenges with commercial software-defined radio platforms”. In: *Communications Magazine, IEEE* 54.1 (Jan. 2016), pp. 59–67. ISSN: 0163-6804. DOI: 10.1109/MCOM.2016.7378427.

[SMS09] Bilal Sadiq, Ritesh Madan, and Ashwin Sampath. “Downlink scheduling for multiclass traffic in LTE”. In: *EURASIP Journal on Wireless Communications and Networking* 2009 (Nov. 2009).

[SN06] M. S. Safadi and D. L. Ndzi. “Digital Hardware Choices For Software Radio (SDR) Baseband Implementation”. In: *2006 2nd International Conference on Information Communication Technologies*. Vol. 2. 2006, pp. 2623–2628. DOI: 10.1109/ICTTA.2006.1684823.

[Sol+07] Stephen Soltesz et al. “Container-based Operating System Virtualization: A Scalable, High-performance Alternative to Hypervisors”. In: *SIGOPS Oper. Syst. Rev.* 41.3 (Mar. 2007), pp. 275–287. ISSN: 0163-5980. DOI: 10.1145/1272998.1273025.

[Spe+04] Q.H. Spencer et al. “An introduction to the multi-user MIMO downlink”. In: *Communications Magazine, IEEE* 42.10 (Oct. 2004), pp. 60–67. ISSN: 0163-6804. DOI: 10.1109/MCOM.2004.1341262.

[Sri+08] Roshni Srinivasan et al. *IEEE 802.16m Evaluation Methodology Document (EMD)*. Tech. rep. IEEE 802.16 Broadband Wireless Access Working Group, July 2008.

[SSM07] G. Sharma, N. B. Shroff, and R. R. Mazumdar. “Joint Congestion Control and Distributed Scheduling for Throughput Guarantees in Wireless Networks”. In: *IEEE INFOCOM 2007 - 26th IEEE International Conference on Computer Communications*. May 2007, pp. 2072–2080. DOI: 10.1109/INFCOM.2007.240.

[Sta+99] J. A. Stankovic et al. “The case for feedback control real-time scheduling”. In: *Real-Time Systems, 1999. Proceedings of the 11th Euromicro Conference on*. 1999, pp. 11–20. DOI: 10.1109/EMRTS.1999.777445.

[Sto05] Alexander L. Stolyar. “On the Asymptotic Optimality of the Gradient Scheduling Algorithm for Multiuser Throughput Allocation”. In: *Oper. Res.* 53.1 (Jan. 2005), pp. 12–25. ISSN: 0030-364X. DOI: 10.1287/opre.1040.0156.

[Stu08] Student. “The Probable Error of a Mean”. In: *Biometrika* 6.1 (1908), pp. 1–25. DOI: 10.1093/biomet/6.1.1.

[Stü11] Gordon L. Stüber. *Principles of Mobile Communication*. Third edition. New York: Springer, 2011. DOI: 10.1007/978-1-4614-0364-7.

[SZ79] Prabhakant Sinha and Andris A. Zoltners. “The Multiple-Choice Knapsack Problem”. In: *Operations Research* 27.3 (1979), pp. 503–515. DOI: 10.1287/opre.27.3.503.

[Tan+11] Kun Tan et al. “Sora: High-performance Software Radio Using General-purpose Multi-core Processors”. In: *Commun. ACM* 54.1 (Jan. 2011), pp. 99–107. ISSN: 0001-0782. DOI: 10.1145/1866739.1866760.

[Tel09] TeliaSonera. *TeliaSonera first in the world with 4G services*. press release, online, archived by http://web.archive.org on 12/17/2009, visited on 12/16/2015. Dec. 2009. URL: http://www.teliasonera.com/press/pressreleases/item.page?prs.itemId=463244.

[Ver84] Sergio Verdú. “Optimum Multi-User Signal Detection”. PhD thesis. Urbana-Champaign, Illinois, 1984.

[Ver86] Sergio Verdú. “Minimum probability of error for asynchronous Gaussian multiple-access channels”. In: *IEEE Transactions on Information Theory* 32.1 (Jan. 1986), pp. 85–96. ISSN: 0018-9448. DOI: 10.1109/TIT.1986.1057121.

[Vri+02] J. De Vriendt et al. “Mobile network evolution: a revolution on the move”. In: *IEEE Communications Magazine* 40.4 (Apr. 2002), pp. 104–111. ISSN: 0163-6804. DOI: 10.1109/35.995858.

[VTR14] Matthew C Valenti, Salvatore Talarico, and Peter Rost. “The role of computational outage in dense cloud-based centralized radio access networks”. In: *Global Communications Conference (GLOBECOM), 2014 IEEE*. IEEE. 2014, pp. 1466–1472.

[Wan+16] X. Wang et al. “Energy-Efficient Virtual Base Station Formation in Optical-Access-Enabled Cloud-RAN”. In: *IEEE Journal on Selected Areas in Communications* PP.99 (2016), pp. 1–1. ISSN: 0733-8716. DOI: 10.1109/JSAC.2016.2520247.

[Wei+12] Jiang Weipeng et al. “Major optimization methods for TD-LTE signal processing based on general purpose processor”. In: *Communications and Networking in China (CHINACOM), 2012 7th International ICST Conference on*. Aug. 2012, pp. 797–801. DOI: 10.1109/ChinaCom.2012.6417593.

[Wer+15] T. Werthmann et al. “Task Assignment Strategies for Pools of Baseband Computation Units in 4G Cellular Networks”. In: *IEEE ICC 2015 - Workshop on Cloud-Processing in Heterogeneous Mobile Communication Networks (IWCPM)*. 2015, pp. 2714–2720.

[Wer15] T. Werthmann. “Approaches to Adaptively Reduce Procedding Effort for LTE Cloud-RAN Systems”. In: *IEEE ICC 2015 - Workshop on Cloud-Processing in Heterogeneous Mobile Communication Networks (IWCPM)*. 2015, pp. 2701–2707.

[WGP13] T. Werthmann, H. Grob-Lipski, and M. Proebster. “Multiplexing Gains Achieved in Pools of Baseband Computation Units in 4G Cellular Networks”. In: *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*. 2013. DOI: 10.1109/PIMRC.2013.6666722.

[WOE05] C. Wengerter, J. Ohlhorst, and A. G. E. von Elbwart. “Fairness and throughput analysis for generalized proportional fair frequency scheduling in OFDMA”. In: *2005 IEEE 61st Vehicular Technology Conference*. Vol. 3. May 2005, 1903–1907 Vol. 3. DOI: 10.1109/VETECS.2005.1543653.

[Wüb+14] D. Wübben et al. “Benefits and Impact of Cloud Computing on 5G Signal Processing: Flexible centralization through cloud-RAN”. In: *Signal Processing Magazine, IEEE* 31.6 (Nov. 2014), pp. 35–44. ISSN: 1053-5888. DOI: 10.1109/MSP.2014.2334952.

[Wüb+15] Dirk Wübben et al. *Final definition and evaluation of PHY layer approaches for RANaaS and joint backhaul-access layer*. Tech. rep. D2.3. FP 7 project iJOIN, Apr. 2015.

[Zak+11] Y. Zaki et al. “Multi-QoS-Aware Fair Scheduling for LTE”. In: *Vehicular Technology Conference (VTC Spring), 2011 IEEE 73rd*. May 2011, pp. 1–5. DOI: 10.1109/VETECS.2011.5956352.

[Zha+16] Z. Zhang et al. “Full-Duplex Wireless Communications: Challenges, Solutions, and Future Research Directions”. In: *Proceedings of the IEEE* 104.7 (July 2016), pp. 1369–1409. ISSN: 0018-9219. DOI: 10.1109/JPROC.2015.2497203.

[Zha95] Hui Zhang. “Service disciplines for guaranteed performance service in packet-switching networks”. In: *Proceedings of the IEEE* 83.10 (Oct. 1995), pp. 1374–1396. ISSN: 0018-9219. DOI: 10.1109/5.469298.

[Zho+16] S. Zhou et al. “Software-defined hyper-cellular architecture for green and elastic wireless access”. In: *Communications Magazine, IEEE* 54.1 (Jan. 2016), pp. 12–19. ISSN: 0163-6804. DOI: 10.1109/MCOM.2016.7378420.

[Zhu+11] ZhenBo Zhu et al. “Virtual Base Station Pool: Towards a Wireless Network Cloud for Radio Access Networks”. In: *Proceedings of the 8th ACM International Conference on Computing Frontiers*. CF ’11. Ischia, Italy: ACM, 2011, 34:1–34:10. ISBN: 978-1-4503-0698-0. DOI: 10.1145/2016604.2016646.

[Zim07] Ernesto Zimmermann. “Complexity Aspects in Near Capacity MIMO Detection Decoding”. Dissertation. Technische Universität Dresden, 2007.