Czech Technical University in Prague

Faculty of Electrical Engineering

Department of Telecommunication Engineering

Allocation of Communication and Computation Resources in Mobile Networks

Doctoral Thesis

Ing. Jan Plachý

Prague, November 2020

Ph.D. Programme: P2612 Electrical Engineering and Information Technology
Branch of study: 2601V013 Telecommunication Engineering

Supervisor: doc. Ing. Zdeněk Bečvář, Ph.D.
Supervisor-Specialist: Dr. Emilio Calvanese Strinati, HDR

I hereby declare I have written this doctoral thesis independently and quoted all the sources of information used in accordance with methodological instructions on ethical principles for writing an academic thesis. Moreover, I state that this thesis has neither been submitted nor accepted for any other degree.

Prague, November, 2020

.....

Signature

Abstract

The convergence of communication and computing in the mobile networks has led to an introduction of the Multi-Access Edge Computing (MEC). The MEC combines communication and computing resources at the edge of the mobile network and provides an option to optimize the mobile network in real-time. This is possible due to close proximity of the computation resources in terms of communication delay, in comparison to the Mobile Cloud Computing (MCC). The optimization of the mobile networks requires information about the mobile network and User Equipment (UE). Such information, however, consumes a significant amount of communication resources. The finite communication resources along with the ever increasing number of the UEs and other devices, such as sensors or vehicles pose an obstacle for collecting the required information. Therefore, it is necessary to provide solutions to enable the collection of the required mobile network information from the UEs for the purposes of the mobile network optimization.

In this thesis, a solution to enable communication of a large number of devices, exploiting Device-to-Device (D2D) communication for data relaying, is proposed. To motivate the UEs to relay data of other UEs, we propose a resource allocation algorithm that leads to a natural cooperation of the UEs. To show, that the relaying is not only beneficial from the perspective of an increased number of UEs, we provide an analysis of the energy consumed by the D2D communication. To further increase the number of the UEs we exploit a recent concept of the flying base stations (FlyBSs), and we develop a joint algorithm for a positioning of the FlyBS and an association of the UEs to increase the UEs satisfaction with the provided data rates.

The MEC can be exploited not only for processing of the collected data to optimize the mobile networks, but also by the mobile users. The mobile users can exploit the MEC for the computation offloading, i.e., transferring the computation from their UEs to the MEC. However, due to the inherent mobility of the UEs, it is necessary to determine communication and computation resource allocation in order to satisfy the UEs requirements. Therefore, we first propose a solution for a selection of the communication path between the UEs and the MEC (communication resource allocation). Then, we also design an algorithm for joint communication and computation resource allocation. The proposed solution then leads to a reduction in the computation offloading delay by tens of percent.

Keywords: Mobile Networks, Multi-access Edge Computing, Offloading, Mobility management, Resource allocation, Real-time

Abstrakt

Konvergence komunikačních a výpočetních technologií vedla k vzniku Multi-Access Edge Computing (MEC). MEC poskytuje výpočetní výkon na tzv. hraně mobilních sítí (základnové stanice, jádro mobilní sítě), který lze využít pro optimalizaci mobilních sítí v reálném čase. Optimalizace v reálném čase je umožněna díky nízkému komunikačnímu zpoždění například v porovnání s Mobile Cloud Computing (MCC).

Pro optimalizaci mobilních sítí je nutný sběr informací o mobilní síti od uživatelských zařízeních. Sběr těchto informací nicméně využívá komunikační prostředky, které jsou využívány i pro přenos uživatelských dat. Zvyšující se počet uživatelských zařízení, senzorů a taktéž komunikace vozidel tvoří překážku pro sběr informací o mobilních sítích z důvodu omezeného množství komunikačních prostředků. Tudíž je nutné navrhnout řešení, která umožní sběr těchto informací pro potřeby optimalizace mobilních sítí.

V této práci je navrženo řešení pro komunikaci vysokého počtu zařízeních, které je postaveno na využití přímé komunikace mezi zařízeními. Uživatelé jsou motivováni k využití přeposílání dat pomocí přímé komunikace díky přidělení více komunikačních prostředků, jenž vede na přirozenou spolupráci uživatelů. Dále, pro ukázání výhod přímé komunikace mezi uživateli, je provedena analýza spotřeby energie přímé komunikace mezi uživateli. Další zvýšení počtu komunikujících zařízení je založeno na využití mobilních létajících základových stanic (FlyBS). Pro nasazení FlyBS je navržen algoritmus, který hledá pozici FlyBS a asociuje uživatele k FlyBS pro zvýšení spokojenosti uživatelů s poskytovanými datovými propustnostmi.

MEC lze využít nejen pro optimalizaci mobilních sítí z pohledu mobilních operátorů, ale taktéž uživateli mobilních sítí. Tito uživatelé mohou využít MEC pro přenos výpočetně náročných úloh z jejich mobilních zařízeních do MEC. Z důvodu mobility uživatel je nutné nalézt vhodné přidělení komunikačních a výpočetních prostředků pro uspokojení požadavků uživatelů. Tudíž je navržen algoritmus pro výběr komunikační cesty mezi uživatelem a MEC. Tento algoritmus je posléze rozšířen o společné přidělování komunikačních a výpočetních prostředků. Toto navržené řešení následně vede ke snížení komunikačního zpoždění o desítky procent.

Klíčová slova: Mobilní sítě, Multi-access Edge Computing, Přesun výpočtů, Správa mobility, Přidělování prostředků, Reálný čas

Acknowledgements

This doctoral thesis is the output of the effort and support of several people to whom I am extremely grateful. First and foremost, I would like to to express my deep gratitude to my supervisor doc. Ing. Zdeněk Bečvář, Ph.D., who has been continuously supporting me and providing mentorship throughout the years. Moreover, I would like to thank him for providing motivation when needed and his patience with me.

I would like to express my gratitude to Dr. Emilio Calvanese Strinati, HDR, who has provided his guidance and knowledge acting as a supervisor-specialist and a precious opportunity to do an internship at CEA-Leti. My thanks go to Dr. Nicola di Pietro, whose technical skills and dedication were invaluable.

Many thanks go to prof. Amir Leshem, with whom I had the privilege to collaborate, as he shared his deep mathematical knowledge not only during my stay at Bar-Ilan University, but throughout the ongoing collaboration. Also, I would like to thank Dr. Syed Mohammad Zafaruddin, with whom I have enjoyed collaboration and our discussions.

I am very grateful for the opportunity to collaborate with prof. Adlen Ksentini and prof. Navid Nikaein and their supervision during my internships at EURECOM by sharing their knowledge with bringing my research towards real world implementation.

Special thanks goes to Dr. Pavel Mach and Dr. Michal Vondra who were always open to discuss my work and provide enlightening ideas and encouragement.

Many thanks go to all colleagues that I have had the opportunity to discuss my research at 5G mobile research lab, CEA-Leti, Bar-Ilan University and EURECOM. This thesis would not exist without the technical support of Czech Technical University in Prague and Department of Telecommunication Engineering, and financial support by all projects that provided funding and made this thesis possible.

Last, but certainly not least, I would like to thank to my family and my fiancée for providing me all the support and motivation needed to finish this thesis.

Glossary

3GPP ETSI 3rd Generation Partnership Project.

4G fourth generation.

5G fifth generation.

AM-AOMDV Adaptive Multi-metric Ad-Hoc On-Demand Multipath Distance Vector.

AOMDV-DPU Ad-hoc On-demand Multipath Distance Vector with Dynamic Path Update.

ARIMA Autoregressive integrated moving average.

ARQ Automatic Repeat reQuest.

AWGN Additive White Gaussian Noise.

BBU Base Band Unit.

BER Bit Error Rate.

BLER Block Error Rate.

BS Base Station.

BSR Buffer Status Report.

C-RAN Cloud-Radio Access Network.

CDF Cumulative Distribution Function.

CLO Cross-layer Optimization.

CPU Central Processing Unit.

CRC Cyclic Redundancy Check.

CSI Channel State Information.

CSMA Carrier Sensing multiple Access.

CTR Clear To Relay.

CTS Clear To Send.

D2D Device-to-Device.

DCCRA Dynamic Communication and Computing Resource Allocation.

DF Decode and Forward.

EAB Extended Access Barring.

eNB 4G base station.

EPA extended pedestrian A.

EPDCCH Enhanced Physical Downlink Control Channel.

FDD Frequency Division Duplex.

FlyBS Flying BS.

GA Genetic algorithm.

GBR Guaranteed Bit Rate.

GFDM Generalized Frequency Division Multiple Access.

gNB 5G base station.

GPU Graphical Processing Unit.

HARQ Hybrid Automatic Repeat reQuest.

HeNB Femto Cell eNB.

HMM Hidden Markov Model.

IoT Internet of Things.

IP Internet Protocol.

IPv4 Internet Protocol version 4.

IPv6 Internet Protocol version 6.

KKT Karush-Kuhn-Tucker.

LOS Line Of Sight.

LTE Long Term Evolution.

LTE-M Long Term Evolution - Machine Type Communication.

LTE-A Long Term Evolution - Advanced.

MAC Medium Access Control.

MCC Mobile Cloud Computing.

MCS Modulation and Coding Scheme.

MDP Markov Devicision Process.

MEC Multi-Access Edge computing.

MEC host MEC host.

MEO Mobile Edge Orchestrator.

MIPS Millions Instructions Per Second.

MLE Maximum Likelihood Estimation.

MTC Machine Type Comunication.

NBS Nash Baragaining Solution.

NC No Compression.

NCL Neighbor Cell List.

O-RAN Open Radio Access Network.

OAI Open Air Interface.

ODSR Opportunistic Device Select Relaying.

OFDMA Orthogonal Frequency Division Multiple Access.

OR Overhead Reduction.

ORS Opportunistic Relay Selection.

OS Operating System.

PDCCH Physical Downlink Control Channel.

PDCP Packet Data Convergence Protocol.

PDF Probability Distribution Function.

PDMPRP Power and Delay-aware Multi-path Routing Protocol.

PDSCH Physical Downlink Shared Channel.

PRACH Physical Random Access Channel.

PSO Particle Swarm Optimization.

PSwH Path Selection with Handover.

QAM Quadrature Amplitude Modulation.

QoE Quality of Experience.

QoS Quality of Service.

QPSK Quadrature Phase Shift Keying.

RAM Random Access Memory.

RAP Random Access Procedure.

RB Resource Block.

RLC Radio Link Control.

ROHC Robust Overhead Compression.

ROHC SO ROHC Second Order.

ROHC FO ROHC First Order.

RSS Received Signal Strength.

RSSI Received Signal Strength Indicator.

RTR Request To Relaying.

RTS Request To Send.

RWS Roulette Wheel Selection.

SBS Static Base Station.

SC-FDMA Single Carrier - Frequency Division Multiple Access.

SCC Small Cell Cloud.

SCeNB Small Cell eNB.

SCgNB Small Cell gNB.

SINR Signal to Interference plus Noise Ratio.

SNR Signal to Noise Ratio.

SO Serving Only.

TB Transmission Block.

TCP Transmission Control Protocol.

TCP/IP Transmission Control Protocol/Internet Protocol.

TDL-A Tapped Delay Line Type A.

TDMA Time Domain Multiple Access.

TPC Transmission Power Control.

TTI Transmission Time Interval.

TTL Time To Live.

UAV Unmanned Aerial Vehicles.

UE User Equipment.

VM Virtual Machine.

VM-OAP VM Online Approximation Placement.

VoIP Voice over IP.

WiFi Wireless Fidelity.

WSN Wireless Sensor Networks.

Contents

Abstract i

Abstrakt iii

Acknowledgments v

List of Figures xvi

List of Tables xix

1 Introduction 1
- 1.1 Motivation . . . 3
- 1.2 Organization of the thesis . . . 4

2 State of the art 5
- 2.1 Mobile networks . . . 5
  - 2.1.1 Collecting mobile network information . . . 6
  - 2.1.2 Device-to-Device communication . . . 9
  - 2.1.3 Beyond 5G communication . . . 10
- 2.2 Multi-Access Edge Computing . . . 12
  - 2.2.1 Communication with the MEC . . . 13
  - 2.2.2 Allocation of computing resources in MEC . . . 15

3 Thesis objectives 18

4 Collecting user and network information 19
- 4.1 Cross-layer optimization of LTE-A signaling . . . 19
  - 4.1.1 Management of the proposed scheme . . . 21
  - 4.1.2 Evaluation of the proposal for collection of data from devices . . . 23
  - 4.1.3 Conclusion . . . 26
- 4.2 Cooperative resource allocation in a relayed communication . . . 26
  - 4.2.1 System Model . . . 27
  - 4.2.2 Problem Formulation . . . 29
  - 4.2.3 Nash Bargaining Solution . . . 29
  - 4.2.4 Simulation Results and Analysis . . . 32
  - 4.2.5 Conclusion . . . 34
- 4.3 Energy Consumption of Opportunistic D2D Relaying Under Lognormal Shadowing . . . 34
  - 4.3.1 System Model . . . 34
  - 4.3.2 ODSR Relaying Scheme . . . 36

4.3.3 Distributed Implementation of ODSR . . . 37
4.3.4 Performance Bounds of ODSR . . . 39
4.3.5 Simulation and Numerical Analysis . . . 45
4.3.6 Conclusion . . . 49
4.4 Increasing number of communicating users beyond 5G . . . 50
4.4.1 System Model and Problem Formulation . . . 50
4.4.2 Proposed Solution . . . 52
4.4.3 Simulation Scenario and Performance Evaluation . . . 59
4.4.4 Conclusion . . . 63

5 Resource allocation in the MEC 64
5.1 Markov Decision Process . . . 65
5.1.1 MDP in communication and computation resource allocation . . . 66
5.2 Selection of communication paths for MEC . . . 67
5.2.1 Path Selection Algorithm . . . 68
5.2.2 System model . . . 68
5.2.3 Path selection exploiting handover . . . 70
5.2.4 Path selection algorithm . . . 71
5.2.5 Implementation aspects . . . 73
5.2.6 Complexity of the path selection algorithm . . . 73
5.2.7 Evaluation Methodology and Scenario . . . 74
5.2.8 Simulation Results . . . 76
5.2.9 Conclusion . . . 83
5.3 Joint computation and communication resource allocation under fixed prediction accuracy . . . 83
5.3.1 System model . . . 84
5.3.2 Dynamic Resource Allocation . . . 86
5.3.3 Performance Evaluation . . . 88
5.3.4 Conclusion . . . 91
5.4 Prediction based communication and computing resource allocation for MEC 91
5.4.1 System Model and Problem Formulation . . . 92
5.4.2 Mobility and Channel Quality Prediction . . . 97
5.4.3 Proposed Dynamic Communication and Computing Resource Allocation Algorithm . . . 103
5.4.4 Simulation Scenario and Models . . . 107
5.4.5 Performance Evaluation and Discussion of Results . . . 109
5.4.6 Conclusion . . . 115

6 Conclusion 118
6.1 Thesis summary . . . 118
6.2 Research contributions . . . 119
6.3 Future research direction . . . 120

References 122

Appendices 139

A Energy consumption of mobile networks **139**
A.1 Uplink transmission power control 139
A.2 Empirical energy consumption model 141

B Approximation of $I_2^{\mathrm{RELAY}}(N, \sigma)$ **143**

C Scaling Law on Energy Consumption **144**

D List of Publications **146**

E List of Projects **148**

F Others **149**

List of Figures

1.1 Resource allocation and FlyBS deployment based on collected UE network information. On the left side, the mobile network information is collected from the UEs at the MEC, where it is processed. The processed mobile network information is then exploited for optimization of BS's resource allocation and deployment of the FlyBS. 2

2.1 Long Term Evolution - Advanced (LTE-A) transmission protocol stack at the device. 6

4.1 Principle of buffering within cluster. 20
4.2 Scenario of the proposed approach for collection of information from devices. 21
4.3 Procedure for the device beginning its transmission of payload in case of (a) no cluster in proximity, (b) joining cluster in proximity, (c) becoming cluster head. 22
4.4 Proposed signaling messages enabling cross-layer optimization of frequent transmission of small payloads. 23
4.5 The number of served devices transmitting frequently small payloads (a) comparison of the proposal and competitive schemes, (b) impact of clustering and buffering. 24
4.6 Overhead ratio (a) by the proposal and competitive schemes, (b) impact of clustering and buffering. 25
4.7 Impact of clustering and buffering on overhead ratio. 26
4.8 System model with device i relaying its data through device j. 27
4.9 Number of transmitted packets ($N_i(s_i)$) with $\gamma_1 = 0$ dB and $\gamma_2 = 30$ dB, (dashed line - Device 1 (source), solid line - Device 2 (relay). 31
4.10 Average energy consumed per transmission ($E_i^{tx}(s_i)$) (a) and total number of transmitted packets ($\sum N_i$) (b). 33
4.11 Jain's fairness index of number of transmitted packets gained by cooperation (a) and Cumulative Distribution Function (CDF) of the gain (b). . . . 33
4.12 D2D relaying in the uplink communication of a single cell network. Devices are inside a shopping mall/university building/ offices and the Base Station (BS) is far away separated by walls. The devices have single antenna while the BS has multiple antennas. 34
4.13 Opportunistic Device Select Relaying (ODSR) for three devices with transmission energy $E_1 < E_2 < E_s$ (a) function $f(E)$ (b) timing diagram and resource block allocation. 36

4.14 Energy consumption performance of opportunistic relaying compared to direct transmission over wireless fading channels for various network scenarios. Different acronyms are UMa: Urban macro, UMi: Urban micro, SC: Street Canyon, NLOS: non-line of sight. 46
4.15 Performance of ODSR comparing with the optimal and no-relaying schemes under 3GPP WINNER II fading channels. 48
4.16 Validation of derived analytical bounds and effect of circuit transmission power on the relaying performance. 49
4.17 Update of Flying BS (FlyBS) position via the proposed algorithm based on Particle Swarm Optimization (PSO). 54
4.18 Update of the FlyBS position by the proposed algorithm based on genetic algorithm. 57
4.19 Simulation area with FlyBSs and Static Base Stations (SBSs) and associated User Equipments (UEs) (association to individual FlyBSs is indicated by colors). 60
4.20 Improvement of network performance in (a) UEs satisfaction and (b) network throughput. 61
4.21 Improvement of network performance in (a) UEs satisfaction and (b) network throughput. 62

5.1 Reward on a chain of a single time step (i.e., $k = 1$). 66
5.2 Chain for calculation of the total reward (for the sake of figure clarity, negative rewards $-t_H$ due to handover between states is not depicted for $t > 1$). 66
5.3 Network topology and definition of parameters required for path selection. 69
5.4 Simulation scenario with example of deployment of buildings, users, Small Cell eNBs (SCeNBs) and 4G base station (eNB) for simulations. 75
5.5 Average time (t_{MEC}) required for transmission of offloaded task with size of (a) 300 kB and (b) 30 MB for DSL backhaul (top subplot) and fiber optic (bottom subplot) backhauls. 77
5.6 Average energy (E) required for transmission of offloaded task with size of (a) 300 kB and (b) 30 MB for DSL backhaul (top subplot) and fiber optic (bottom subplot) backhauls. 78
5.7 Ratio of users satisfied with experienced delay, R_S, for DSL backhaul for offloaded task of 300 kB (a) and 30 MB (b). 79
5.8 Ratio of users satisfied with experienced delay, R_S, for optical fiber backhaul, for offloaded task of 300 kB (a) and 30 MB (b). 79
5.9 Mean number of the offloaded tasks, μ_T, transmitted over DSL backhaul for the offloaded task size of 300 kB (a) and 30 MB (b). 80
5.10 Mean number of the offloaded tasks, μ_T, transmitted over optical fiber backhaul for the offloaded task size of 300 kB (a) and 30 MB (b). 81
5.11 Ratio of additional handovers generated by the Path Selection with Handover (PSwH) algorithm, R_H, with respect to the Serving Only (SO) for offloaded tasks of 300 kB (a) and 30 MB (b). 82
5.12 System model. 85
5.13 Offloading times required to offload, compute and collect results of the offloaded task, Average time (a) and CDF of time (b). 90

5.14 Energy consumption of UE communication, Average energy (a) and CDF of energy consumption (b). 91
5.15 MEC system model with one mobile UE. 94
5.16 Example of UE mobility prediction with one degree of mobility freedom following a curved street with known street center positions. 99
5.17 Example of the UE with multiple degrees of mobility freedom, arriving from angle v_3 (red dashed line) with multiple options for the departure angle w (solid lines). 101
5.18 Cooperation of the proposed Dynamic Communication and Computing Resource Allocation (DCCRA) algorithms. 104
5.19 Example of Virtual Machine (VM) placement by Algorithm 7 for three 5G base stations (gNBs) (number of rows in each table in the middle part of the figure) over five time instants (columns in each table in the middle part of the figure). For each departure angle, represented by individual table, a sequence (row) of the gNBs maximizing α is chosen and then exploited to determine t_S and t_E for each gNB. 106
5.20 Simulation model with deployment of gNBs and SCgNBs. 109
5.21 Mean times required to offload, compute, and collect results of the offloaded task for 30 UEs with $L_{\mathrm{O}} = L_{\mathrm{C}} = 20$ kB (a), $L_{\mathrm{O}} = L_{\mathrm{C}} = 200$ kB (b), and $L_{\mathrm{O}} = L_{\mathrm{C}} = 2000$ kB (c). 109
5.22 Mean offloading energy for 30 UEs with $L_{\mathrm{O}} = L_{\mathrm{C}} = 20$ kB (a), $L_{\mathrm{O}} = L_{\mathrm{C}}$ = 200 kB (b), and $L_{\mathrm{O}} = L_{\mathrm{C}} = 2000$ kB (c). 110
5.23 Mean times required to offload, compute, and collect results of the offloaded for 60 UEs with $L_{\mathrm{O}} = L_{\mathrm{C}} = 20$ kB (a), $L_{\mathrm{O}} = L_{\mathrm{C}} = 200$ kB (b), and $L_{\mathrm{O}} = L_{\mathrm{C}}$ = 2000 kB (c). 111
5.24 Mean offloading energy for 60 UEs with $L_{\mathrm{O}} = L_{\mathrm{C}} = 20$ kB (a), $L_{\mathrm{O}} = L_{\mathrm{C}}$ = 200 kB (b), and $L_{\mathrm{O}} = L_{\mathrm{C}} = 2000$ kB (c). 111
5.25 Mean times required to offload, compute, and collect results of the offloaded for 90 UEs with $L_{\mathrm{O}} = L_{\mathrm{C}} = 20$ kB (a), $L_{\mathrm{O}} = L_{\mathrm{C}} = 200$ kB (b), and $L_{\mathrm{O}} = L_{\mathrm{C}}$ = 2000 kB (c). 113
5.26 Mean offloading energy for 90 UEs with $L_{\mathrm{O}} = L_{\mathrm{C}} = 20$ kB (a), $L_{\mathrm{O}} = L_{\mathrm{C}}$ = 200 kB (b), and $L_{\mathrm{O}} = L_{\mathrm{C}} = 2000$ kB (c). 113
5.27 Mean amount of data transmitted per backhaul per task for $L_{\mathrm{O}} = L_{\mathrm{C}} =$ 20 kB (a), $L_{\mathrm{O}} = L_{\mathrm{C}} = 200$ kB (b), and $L_{\mathrm{O}} = L_{\mathrm{C}} = 2000$ kB (c), solid line represents 30 UEs, dashed 60 UEs and dotted 90 UEs. 114
5.28 Number of pre-allocated VMs for the proposal during the simulation run. for $L_{\mathrm{O}} = L_{\mathrm{C}} = 20$ kB (a), $L_{\mathrm{O}} = L_{\mathrm{C}} = 200$ kB (b), and L_{O}=$L_{\mathrm{C}} = 2000$ kB (c), solid line represents 30 UEs, dashed 60 UEs and dotted 90 UEs. 114

A.1 Example of tradeoff between energy and time consumed by transmission of 100 kB using 10 Resource Blocks (RBs) with path loss of 80 dB. 140

List of Tables

4.1 Average energy consumption (in μJ) of various overheads obtained using simulation under 3GPP model. 47
4.2 Simulation parameters. 59

5.1 Simulation parameters 76
5.2 Parameters of backhaul models. 76
5.3 Summarized improvement (green color) in performance metrics introduced by the PSwH comparing to the SO for proper values of γ. 83
5.4 Simulation parameters 88
5.5 Simulation parameters. 108

Chapter 1

Introduction

Mobile networks of the fifth generation (5G) are foreseen to enable communication of a huge number of connected devices (e.g., smartphones, tablets, sensors, or machines). The 5G should enable communication of trillions of devices in 2020 [1]. The increase in the number of devices is based on an introduction of Internet of Things (IoT) and Machine Type Comunication (MTC), and on continuous increase in the number of conventional user devices, such as smartphones or tablets. It is expected that a single base station will serve between ten and hundred thousand MTC devices and thousands of conventional User Equipments (UEs) [2]. Also, at the same time, the number of network parameters to be configured, in order to optimize performance of the network, is expected to increase from 1500 in fourth generation (4G) to 2000 in 5G [3]. This motivates self-optimization of the mobile networks [4]. For an efficient self-optimization of the mobile networks, a huge amount of information should be collected and further processed. Such information may range from radio parameters, such as signal quality, to information related to positions or mobility of the devices (mobile phones, tablets, sensors, etc.). A common indicator of all expected information is relatively small volume of data collected from many devices with a relatively high frequency. The communication of a large number of devices can be tackled by exploiting Device-to-Device (D2D) for data relaying [5, 6]. However, the D2D introduces several challenges, such as resource allocation [7, 8] or power allocation [9]. Another option is to deploy a base stations (BSs) on an Unmanned Aerial Vehicless (UAVs) leading to an introduction of a FlyBS [10]. Nevertheless, the mobility of the FlyBS introduces additional challenges, such as positioning of the FlyBSs [10] or an association of the UEs to the FlyBSs [11].

Apart from enabling the communication of a huge number of devices, the 5G mobile networks disrupt the current separation of the communication and computation resources by converging them together [2]. This convergence leads to an introduction of Multi-Access Edge computing (MEC) [12], where computing resources are distributed at the edge of the mobile network, i.e., BSs. The MEC provides computation resources for control and management of the mobile networks, and enables real-time network self-optimization and cloud computing at the edge of mobile network. The computation resources are generally provided either as Virtual Machines (VMs) [13] or containers [14].

The VMs and containers enable management of the computation resources of multiple users, separate their data and provide the computation power based on the user's demand.

Compared to a conventional Mobile Cloud Computing (MCC) [15], the MEC significantly reduces communication delay for computing the tasks offloaded from the UEs to the cloud and reduces communication load of the backhaul due to its location at the edge of the mobile networks. The architecture of the MEC enables the MEC host (MEC host), providing the computation resources, to be deployed at one of the possible places, based on its location in relation to the mobile network [16]. One of the locations is "bump in the wire", where the MEC host is placed on a BS, represented by either an eNB in case of 4G or a gNB in case of 5G mobile networks. In the thesis, it is assumed that the MEC host is deployed as a "bump in the wire", which is similar to a concept of Small Cell Cloud (SCC) [17–19]. Furthermore, primary focus is given to offloading functionalities of the MEC, enabling offloading of user's applications from the UE, such as facial/object recognition, video/speech processing, mobile gaming, augmented or virtual reality, etc. The offloaded application processes data of the user's application, denoted as the offloaded task. The concept of the mobile network optimization is shown in Figure 1.1, where, on the left side, information about the mobile network from the UEs is collected and, on the right side, exploited for the mobile network optimization (resource allocation of the BS and FlyBS deployment). Furthermore, to illustrate the computation offloading, the mobile UE offloads its computation task to the BS on the left, and, after moving close to the right BS, collects the processed task.

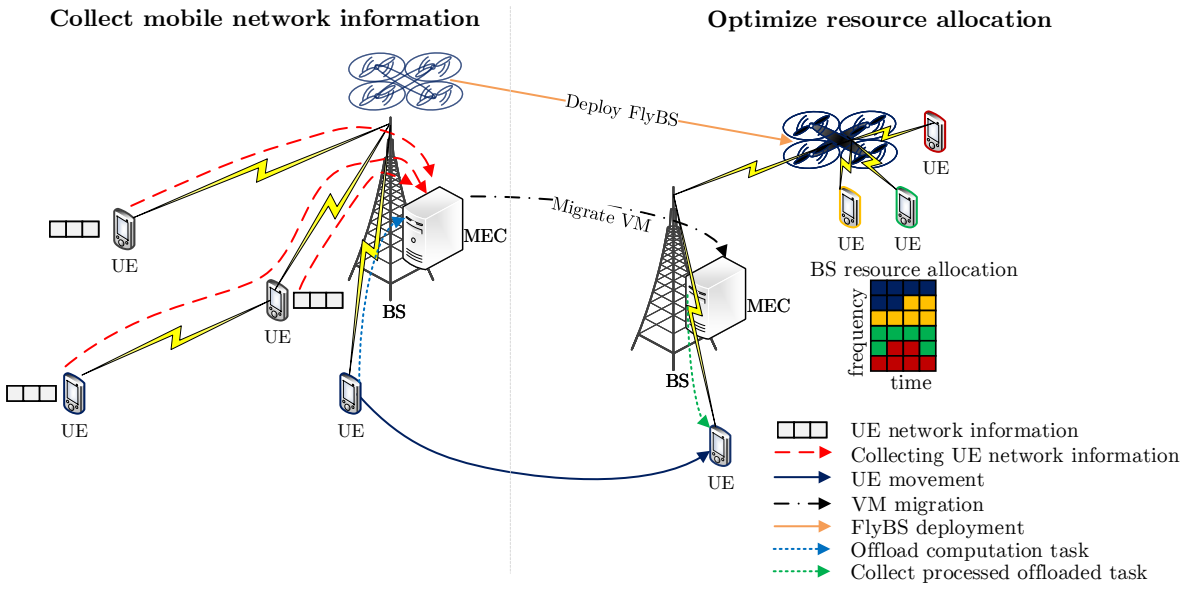


Figure 1.1. Resource allocation and FlyBS deployment based on collected UE network information. On the left side, the mobile network information is collected from the UEs at the MEC, where it is processed. The processed mobile network information is then exploited for optimization of BS's resource allocation and deployment of the FlyBS.

In the rest of this chapter, we provide the motivation for the research work carried

out in the thesis and provide an outline of the thesis.

1.1 Motivation

The upcoming mobile networks improve the Quality of Service (QoS) for the mobile users by increasing their capacity, decreasing communication delay and by providing novel functionalities, such as MEC. However, this comes at a cost of an increased complexity that should be tackled in order to maximize the efficiency of the mobile networks in terms of communication parameters, such as the data rate or delay, and also the energy efficiency, represented by the energy consumed by the UEs communication [3]. To maximize the efficiency of the mobile networks while tackling the problem of the increased complexity, the mobile networks should be able to optimize itself in a real-time. This leads to the self optimizing networks, that collect the information about the mobile network and the UEs, and optimize the communication parameters [4]. However, with the ever increasing number of the UEs and machines (sensors, vehicles, etc.), it is necessary to provide solutions to collect the mobile network information [1]. Moreover, the mobile network information have to be collected at some element in the mobile network and processed before any optimization of the mobile network can be carried out.

The collected data can be stored and processed in the cloud [15]. However, the problem of the real-time optimization persists, as the cloud computing is generally located far away from the BSs and the UEs that are served by these BSs. Thus, it is beneficial to exploit the MEC that provides storage and computation capabilities at the edge of the mobile network [12]. Based on the collected and processed mobile network information from the UEs, the mobile network optimizes its parameters to improve the QoS for the UEs.

One of the recent issues to tackle, is an allocation of the communication and computation resources for the UEs exploiting the MEC [20]. The communication is carried out by the mobile networks, but it is still necessary to lower the communication delay to provide a seamless service. The handover, i.e., changing the serving BS, introduces a significant communication delay, that affect the QoS, and should be considered in the communication resource allocation. The computation resources are provided either in a form of VMs or containers. For the static UEs a solution can be found and kept, i.e., VM or container location is unchanged, until the UE stops exploiting the MEC [21]. However, in case of the mobile UE the problem becomes complex to solve. Therefore, dynamic solutions that optimize the mobile network in real-time are necessary. The dynamic solutions can exploit either a migration of VM [22] or container to another BS or staring a new VM [23] or container [14] at another BS. Nevertheless, moving the computation resources is a time consuming process and leads to a degradation of the QoS. Thus, the moving of the computation resources should be considered in allocation of the computation resources to the mobile UEs.

1.2 Organization of the thesis

In this section, organization of the thesis is described.

Chapter 2 - State of the art provides a deep insight into the collection of the mobile network information from the UEs and resource allocation in the MEC. Section 2.1 provides information about current limits of the mobile networks in terms of maximal number of connectedUEs, exploitation of the D2D for data relaying, and deployment of the FlyBSs. Section 2.2 describes the MEC and existing solutions for the communication and computation resource allocation.

Chapter 3 - Thesis objectives defines thesis objectives based on the state of the art and the motivation.

Chapter 4 - Collecting user and network information consists of the proposed solutions for increasing the number of communicating UEs to collect the network information in Section 4.1. Then cooperative communication resource allocation based on the Nash Baragaining Solution (NBS) is described in Section 4.2, followed by a energy consumption analysis of the D2D relaying in Section 4.3. Section 4.4 deals with a solution of for a deployment of the FlyBSs to improve the QoS of the UEs.

Chapter 5 - Resource allocation in the MEC focuses on exploitation of the MEC for computation offloading of the UEs, and an allocation of the communication and computation resources. First, in Section 5.1, the Markov Devicision Process (MDP) and how it is exploited in the proposed resource allocation algorithms is outlined. Then, in Section 5.2 the algorithm for a selection of the communication path is described. In Section 5.3 and Section 5.4, the proposed joint communication and computation resource allocation with fixed mobility prediction accuracy and unknown mobility are described.

Chapter 6 - Conclusion provides a summary of all achieved results and outlines the future research directions.

Chapter 2

State of the art

In this section, we describe communication limits of the mobile networks in terms of number of UEs communicating and collecting mobile network information from the UEs. Then, the MEC architecture and challenges related to the management of the UEs' mobility in the MEC are explained.

2.1 Mobile networks

Collection of the mobile network information from the UEs is one of the key requirements for self-optimization of the mobile network. However, in the 4G mobile networks (e.g., LTE-A) it is not possible to collect the required information from the UEs, as its amount is very large and consists of a small sized data. This is due to the design of current 4G mobile networks, which are designed to support high speed data transmissions of a large payload (e.g., video, file sharing, etc.). Nevertheless, when it is required to serve a huge number of the UEs sending or receiving a relatively small volumes of data, performance of the LTE-A mobile networks becomes significantly degraded [24, 25]. In current mobile networks, the majority of traffic is being transmitted in downlink rather than in uplink [26]. However, collection of data from the UEs can turn this situation over and, as expected, increase uplink utilization.

Moreover, as the ETSI 3rd Generation Partnership Project (3GPP) mobile networks should be able to self-optimize in real-time [4], the question is how to process the collected mobile network information to optimize th mobile network in a real-time. The collected mobile network information can be processed either centrally [15] or by distributed computing resources deployed closer to the edge of the mobile networks (e.g., the MEC [12]). For both centralized and distributed processing of the mobile network information, load of the radio channels will increase significantly due to a need for gathering of small payloads with significant overhead from a large number of UEs [27].

To improve the performance of the mobile networks a relayed communication based on the D2D can be exploited. It provides an option to a UE to overcome low quality communication channel to its serving BS by relaying its communication over neighboring UEs. Moreover, it can be exploited for collecting of the mobile network information from

the UEs. Another way how to improve the mobile network performance is to deploy the UAVs acting as FlyBSs. The FlyBSs improve the channel quality of the UEs by increasing the probability of Line Of Sight (LOS) communication that leads to an improved channel quality [28].

In the following subsections, we describe limitations of collecting the mobile network information from the UEs, followed by incorporation of the D2D communication and the FlyBSs in the mobile networks.

2.1.1 Collecting mobile network information

Transmission of small payloads from a high number of densely spread devices (UEs, sensors, vehicles, etc.) is currently an issue for the LTE-A based mobile networks as the transmission protocol stack, as shown in Figure 2.1, is not prepared to handle it [29]. Limitations are seen in maximal number of devices to be scheduled within a single subframe [24] as well as due to the collisions of devices trying to connect to the mobile network [25]. Also, an important problem comes from transmission of significant overhead at all layers of the protocol stack, i.e., Physical layer, Control layers and Transmission Control Protocol/Internet Protocol (TCP/IP) layer, as shown in Figure 2.1. The overhead and limitations from each protocol stack layers are described in the following subsections.

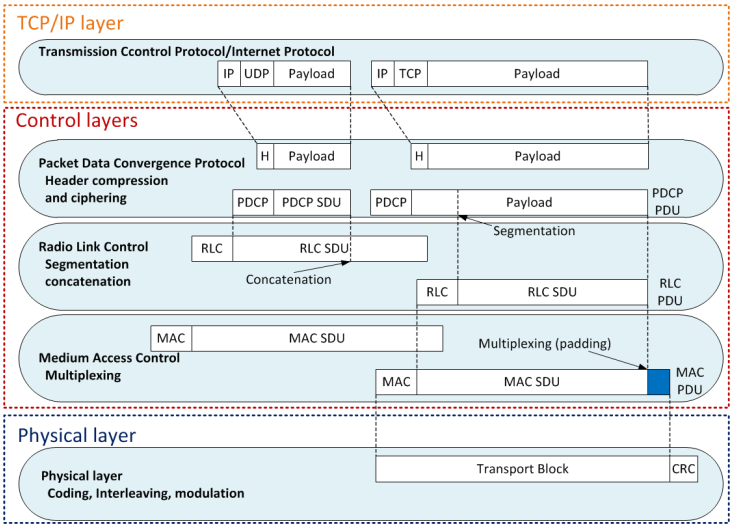


Figure 2.1. LTE-A transmission protocol stack at the device.

Physical layer

The first limitation on the maximal number of served devices is implied by the physical layer as it defines the amount of bits the devices can transmit within a time period. On the radio, LTE-A defines a frame with a duration of 10ms. The frame is divided into 10 subframes (each with duration of 1ms). The minimum amount of bits allocated to one device is defined by the minimal amount of resources per subframe allocated to the device. Each subframe is composed of two RBs in time domain. In LTE-A, at least two time consecutive RBs (i.e., subframe) must be allocated to the device [30]. Depending

on used Modulation and Coding Scheme (MCS), the device can send between 32 and 616 bits per subframe in one Transmission Block (TB) using Quadrature Phase Shift Keying (QPSK) and 64 Quadrature Amplitude Modulation (QAM), respectively (see [31] for more details on relation between TB size, number of RBs, and MCS). The TB contains the device's payload and headers added by all layers as shown in Figure 2.1. If the device is willing to send less bits than the amount, which can be transmitted in RBs allocated to the device, the MCS for transmission can be lowered to reduce transmission error rate. However, from spectral efficiency point of view, this approach is very inefficient and leads to wasting of radio resources. To enable transmission of less than two RBs, Long Term Evolution - Machine Type Communication (LTE-M) has been proposed. The LTE-M aims on the MTC and reduces transmission bandwidth to enable use of single RB (i.e., a half of the subframe) [32, 33]. However, this requires to use Generalized Frequency Division Multiple Access (GFDM) (see more details in [34]) for multiplexing instead of Single Carrier - Frequency Division Multiple Access (SC-FDMA), which is defined for uplink in LTE-A.

Next limit for the uplink transmission originates from the Random Access Procedure (RAP), which serves for initial communication with the eNB. The procedure consists of randomly selected preamble sent by a device to identify itself over the Physical Random Access Channel (PRACH). As there is a limited number of preambles to distinguish each device, collision may occur at the PRACH [25]. Collision probability can be reduced by use of Extended Access Barring (EAB), which is barring communication of low-priority devices. Results from analysis in [35] show that the EAB decreases collision probability but at the cost of increased delivery delay. Different way to avoid the collisions is to reduce the number of the RAP by buffering of several payloads from the device [36]. Instead of starting the RAP each time the device has a payload to send, the RAP is used once for transmission of multiple buffered payloads. Nevertheless, buffering must respect delay constraints of each type of payload. Another way to overcome the PRACH limitation is to dynamically allocate more resources to the PRACH [37]. On one hand, it enables more devices to initiate communication via RAP. On the other hand, this consumes resources commonly allocated to the device for communication. Consequently, a higher number of the devices can be able to access radio resources, but these resources might not be available to all of them in required quantity. Finally, the number of devices being able to transmit required payload by single eNB might not be increased sufficiently.

Control layers

After successful RAP, the device has to be scheduled in order to transmit its payload. In general, three options of scheduling are known: persistent, semi-persistent, or non-persistent (or dynamic) [38]. The persistent scheduling allocates resources to the device for a given period defined by the number of Transmission Time Interval (TTI). An advantage is that the resources are allocated once for the period of multiple TTI, which leads to transmission of less signaling overhead. However, if the device is not transmitting any data in some TTIs, its RBs cannot be reallocated to another device and these RBs are wasted.

In case of non-persistent scheduling, the number of devices scheduled in one TTI is limited to 10 due to the limitations imposed by the Uplink grant (UL grant) information carried inPhysical Downlink Control Channel (PDCCH) [24]. The limit of 10 devices is due to the number of resources available for the UL grant. To overcome the limit of 10 devices scheduled by the PDCCH, 3GPP Release 11 introduces Enhanced Physical Downlink Control Channel (EPDCCH) [31,39]. The EPDCCH overcomes the limitation of PDCCH by utilizing more resources from Physical Downlink Shared Channel (PDSCH) (used for user data transmission in downlink) for the purposes of the UL grant transmission. The last type of scheduling is semi-persistent. The semi-persistent scheduling periodically allocates RBs for the device. This is used for Voice over IP (VoIP) as it has deterministic payload size and regular periodicity of transmission [40]. Since the payloads in semi-persistent scheduling are periodical and of defined size, we can utilize this scheduling to overcome limitations of PRACH and PDCCH. However, this requires constant size of data transmitted by the device. If data is not of constant size, a part of resources has to be reserved for the dynamic scheduling to accommodate bits not fitting to the resources allocated by semi-persistent scheduling. To schedule adequate number of resources to each device, the eNB can exploit knowledge of the device's buffer status (how many bytes are ready to be sent by the device) obtained via Buffer Status Report (BSR) message [41]. The BSR is send by the device in Logical Channel ID field within Medium Access Control (MAC) header. The BSR is send if: i) new data is in buffer of the device, ii) the eNB requests BSR, or iii) there would be more padding bits in MAC header than the length of the BSR itself. However, the BSR can report only specific ranges of payload sizes in the buffer as specified by 3GPP [42]. If less than 10 bytes are in the buffer of the device, the BSR informs the eNB that the device has between 1 and 10 bytes of payload in the buffer. Consequently, 10 bytes are allocated to the device. However, 10 bytes allocated to the device can be more than what the device actually requires. This, then, leads to wasting of resources and less devices can be served.

TCP/IP layer

At the TCP/IP layer, the payload of device is encapsulated in the Transmission Control Protocol (TCP) and Internet Protocol (IP) to enable communication through the IP based networks. The TCP header is typically of 20 bytes, whereas a size of the IP header depends on version of the IP used for communication. For the Internet Protocol version 4 (IPv4) and Internet Protocol version 6 (IPv6) headers, 20 and 40 bytes are required, respectively. The TCP/IP header can be compressed by the Robust Overhead Compression (ROHC) [43]. The ROHC avoids transmission of full TCP/IP headers if the device sends multiple packets to the same destination. Note that the ROHC can be used only for point-to-point connections. The ROHC sends only the dynamically changing parts of the TCP/IP headers to reduce the overhead. The ROHC can work in three modes: Unidirectional, Bidirectional Optimistic, and Reliable. In case of uplink connection without the need of correct delivery acknowledgement, the ROHC works in Unidirectional mode. Other modes utilize downlink for transmission of additional signaling (acknowledgement

of ROHC signaling), thus, we focus on Unidirectional mode only. In this mode, the device (after a given number of packets in a given state) switches periodically between one of the three ROHC states [43]: the Initialization and Refresh, the First Order and the Second Order. In each state, the ROHC sends different signaling with different size in order to provide sufficient robustness. This periodic switching between the states causes problems to scheduling as not all resources can be scheduled using semi-persistent scheduling. Part of the resources has to be reserved for non-persistent scheduling to send the bits not fitting to the resources allocated by semi-persistent scheduling.

This problem can be partly solved by ROHC [43], which reduces TCP/IP overhead. Furthermore, the overhead can be reduced by buffering of several payloads to send them at once [36]. Other possible solution is to cluster nearby users and send their payload merged into a single packet with less overhead as expected in, for example, wireless sensor networks [44].

With the introduction of the MTC and IoT, LTE-M has been proposed to solve the problem with transmission of small payloads. The LTE-M aims on the MTC and reduces transmission bandwidth to enable use of a single RB (i.e., smallest allocable radio resource unit) [32, 33]. However, this requires to use GFDM (see more details in [34]) for multiplexing instead of SC-FDMA, which is defined for uplink in LTE-A. Thus making it backward incompatible with the LTE-A due to use of GFDM.

2.1.2 Device-to-Device communication

In the upcoming mobile networks, BSs are expected to serve a mix of common human traffic and MTC. The number of connected devices, generating both human as well as MTC traffic, is exponentially increasing. This motivates development of efficient strategies handling the traffic generated by these devices, such as exploitation of the D2D communication. The D2D communication enables direct communication between the devices, either in an assisted mode, where the serving BS determines on which resources to communicate and what transmission power should be used, or in an unassisted mode, where the devices have to determine communication parameters themselves. In [5,6], it is shown that relaying of data using the D2D communication to the BS is a feasible solution enabling communication of a massive amount of devices. At the same time, the D2D also reduces energy consumption of the devices [45] but the relaying devices should be motivated [46], as their energy is being consumed. Furthermore, the D2D relaying provides a solution to deal with the channel fading, e.g., shadowing caused by the environment [47–53]. The benefits of the D2D for the devices are further described in [45, 46].

In the D2D communication, most of existing works focus primarily on maximization of data rates [54]. Nevertheless, an energy consumption of communication and relaying is a crucial factor as it impacts the devices' battery lifetime. An experimental analysis of an out-band D2D relaying (communication on a frequency different from the frequency the BS operates on) scheme is presented in [55] to integrate the D2D communications in the mobile network. The authors in [56] derived a geometrical zone for an energy

efficient D2D relaying. In [57], a network-assisted opportunistic D2D clustering has been analyzed in terms of throughput, energy efficiency, and fairness under Rayleigh fading channel models. Considering the D2D fading links as Rician distributed, power control methods have been devised to optimize the power consumption and throughput of the mobile networks [58–60]. A joint optimization of uplink subcarrier assignment and power allocation in the D2D underlying mobile networks is investigated to minimize the energy cost of all users [61]. Recently, the authors in [62] model an energy consumption for the Wireless Fidelity (WiFi) direct which enables the D2D communications between devices in proximity. The performance of the relay-assisted mobile networks under fading channels has been studied for various parameters such as outage probability, throughput, Signal to Noise Ratio (SNR), and Bit Error Rate (BER), but the issue of the energy consumption has not been yet considered. Even for the conventional performance parameters, most of the works ignore the large scale shadowing effect and focus on the short term fading. The shadow fading is modeled using the lognormal distribution which is generally considered harder for performance analysis comparing to the short term fading models. In [63], the authors analyzed the average SNR performance of the opportunistic relaying techniques under large scale channel effects.

The battery lifetime extension motivates cooperation of the devices. The cooperation of devices exploiting D2D can be achieved via Game theory, as shown in [64] for resource allocation, or in [65] for power allocation and channel reuse for D2D communication. A natural solution based on Game theory for the problem of cooperation among devices is the NBS. The NBS has been used to encourage cooperation in other setups of wireless communications, for example, in problems of allocating spectrum over frequency selective channels in Orthogonal Frequency Division Multiple Access (OFDMA) systems [7, 8], for device association to the BSs [66], or for power and bandwidth allocation to the devices [9]. In [67] the authors propose a NBS to maximize data rates of devices via channel assignment and power allocation. However, the authors do not consider communication energy consumption. In [68] the authors consider the NBS for energy efficient resource allocation for D2D relaying but only for two D2D pairs of devices, where each D2D pair provides relaying for the other D2D pair. Thus, making it applicable only in a case of mutual benefit of D2D pairs and impractical for an arbitrary number of D2D pairs acting as relays. Furthermore, the authors do not provide a closed form bargaining solution, but instead formulate the NBS and then solve numerically.

2.1.3 Beyond 5G communication

The increasing requirements of mobile users on wireless communication call for an ultra-dense deployment of BSs [69]. However, the ultra-dense deployment is not always reasonable from an economic point of view. An example of an uneconomical case is an event in which people stay for a short period of time (a few hours) and then move away. This situation occurs mostly during large social events, such as live concerts or sport activities. In these cases, exploitation of a BS mounted on an UAV, also known

as a FlyBS, is beneficial [10]. Nevertheless, a deployment of the FlyBSs brings many challenges, such as finding an optimal position of each FlyBS, properly associating the UEs to the FlyBSs [11], mitigating interference [70, 71], and deciding how many FlyBSs should be deployed in a given area [28, 72].

The current research addressing the problem of the FlyBS positioning can be divided into works dealing with a single FlyBS or those concerning multiple FlyBSs. The positioning of a single FlyBS is considered in, e.g., [10, 73–76]. The objective of the authors in [73] is to find the optimal position of the FlyBS to maximize the data rate of the UE. The FlyBS is seen as a relay between the SBS and the UE. The goal is, then, achieved by a designed algorithm, which finds a position of the FlyBS so that line of sight (LOS) communication takes place on both BS-FlyBS and FlyBS-UE links. The optimal 3D placement of the FlyBS to provide coverage to all UEs is proposed in [74]. The authors derive an optimal position of the FlyBS based on a geometry and a path loss model for a single FlyBS. Similarly, in [75], the authors propose an optimal 3D placement algorithm for a single FlyBS. The algorithm performs an exhaustive search over a closed region to find the 3D position of one FlyBS. The authors assume heterogeneous QoS requirements, represented by the SNR. The positioning is investigated also in [10], where the authors confirm improvement in channel quality, throughput, and energy efficiency by the FlyBS positioned according to the UEs' requirements in a scenario with the UEs moving in a crowd. A joint optimization of the FlyBS's position, bandwidth allocation, transmission power, and transmission rate is proposed in [76]. The authors transform a non-convex optimization problem into a monotonic optimization and solve the problem via the polyblock algorithm [77]. A drawback of all above-mentioned works ([10,73–76]) is that these assume just one FlyBS, and their extension toward multiple FlyBSs is neither easy nor straightforward.

The positioning of multiple FlyBSs is proposed, for example, in [78], where the objective is to provide SNR above a predefined threshold. The authors solve the positioning via linear programming. The drawback of the proposed approach is that it does not take interference into account. Thus, without managing the problem of interference (e.g., by interference alignment technique [79]), such simplification leads to a general coverage optimization problem (as addressed in [80–85]). The interference among individual FlyBSs is assumed in [86], where the authors propose an algorithm for optimization of the FlyBSs' 3D positions. Thus, when compared to the previous papers, the positioning of the FlyBSs is based on Signal to Interference plus Noise Ratio (SINR) of the UEs instead of SNR. The authors focus on the stochastic geometry approach; however, actual data requirements of the UEs are not considered.

Furthermore, evolutionary algorithms [87] can be exploited for the positioning of FlyBSs, as considered in [88–90]. To be more specific, the authors in [88] adopt PSO [91] for the positioning. The PSO finds an optimal solution via an evolutionary process inspired by nature, which acts similar to the flocking of birds or swarms of insects. The authors in [88] propose an algorithm for the positioning of the FlyBSs to provide coverage to all UEs, considering connection quality of the FlyBSs' backhaul. The authors focus on pro-

visioning of a certain level of SINR for the UEs, but they do not consider the allocation of bandwidth to the UEs. The PSO is also exploited in [89], where the authors find an optimal placement of the FlyBSs to satisfy the UEs' required SINR with a minimum number of the deployed FlyBSs. However, the authors do not tackle the problem of bandwidth allocation, which is a critical factor to satisfy the data rates required by the UEs. Furthermore, the authors define SINR requirements as the ratio of the area covered by two or more FlyBSs to the sum of the FlyBSs' coverage areas. Such definition leads to coverage optimization instead of UE data rate satisfaction. Another evolutionary algorithm, the Genetic algorithm (GA) [87], is used in [90] for an optimization of the trajectories of the UAVs. The authors show that the proposed solution based on the GA is efficient and can be run on a Graphical Processing Unit (GPU), exploiting parallel architecture of the GPU. However, the paper does not consider communication of the UAVs with the UEs.

Regarding the problem of the UEs' association to the FlyBSs, the authors in [92] propose two algorithms to partition an area served by the FlyBSs via association of the UEs to the FlyBSs. The objective of the first algorithm, based on optimal transport theory, is to maximize a fairness of the UEs' data rates under a hovering time constraint. The purpose of the second iterative algorithm is to determine a minimal hovering time to satisfy the UEs' data rate requirements. Nevertheless, the positioning of the FlyBSs to improve the UEs' satisfaction is not addressed in [92].

The main disadvantage of all above-mentioned works is that these try to either solely optimize positioning of the FlyBSs or association of the UEs. Nonetheless, the positioning of the FlyBSs and the association of the UEs should be optimized jointly, as these two challenges are closely related. The joint positioning of the FlyBSs and association of the UEs is addressed in [93], where the problem is translated into a clustering problem. This problem is solved by the k-means algorithm, which determines positions of the FlyBSs and associations of the UEs, respectively. The k-means clusters the UEs and associates them to the FlyBSs based on the Euclidean distance. However, the k-means does not incorporate any information regarding the communication channel, which is of paramount importance for the deployment of the FlyBSs.

2.2 Multi-Access Edge Computing

To satisfy high demands of the UEs on computation, the computing power can be distributed over multiple MEC hosts to form computing clusters [19].

The processing of the UEs' applications in the MEC exploits virtualized computing resources in a form of either containers [14] or Virtual Machines (VMs) [13]. Both the VMs and the containers exploit physical computing resources, such as computing time of Central Processing Unit (CPU) or Random Access Memory (RAM), and virtualize them. In the case of containers, the virtualized resources are provided by the containers sharing host's Operating System (OS)). In the case of VMs, the physical computing resources are virtualized by a hypervisor running either on the host OS or directly on the host hardware. Therefore, the VMs and the host OS are completely isolated from each other.

This isolation provides a certain level of security, but the security comes at the cost of an additional overhead [94]. The overhead affects a performance and a startup time of the VMs and makes the VMs less efficient comparing to the containers [95, 96].

An application to be processed in the MEC, denoted as the offloaded application in this thesis, is run on the virtualized resources (VMs or containers). The UE sends data to be processed (denoted as the offloaded task) to the offloaded application in the MEC. For example, the offloaded task by an augmented reality application contains information on the UE's position, its cone of vision, etc. [97]. With focus on offloading of real-time applications, the VM assigned to the UE should be ready when the computing task is being offloaded [98]. Otherwise, delay due to creating and starting VM would make such service unusable.

The offloading process is not always feasible or beneficial due to latency constraints, computation complexity, energy consumption, or memory requirements [99]. Thus, the offloading process is preceded by a decision whether to offload or not [17, 99, 100]. If this decision is positive, the whole offloading process goes through the following stages: i) transmission of the offloaded task to the MEC server where the computing resources are allocated (MEC host), ii) processing of the offloaded task by the offloaded application running over the allocated computing resources, and iii) transmission of the computing results back to the UE. Each stage introduces a delay contributing to the overall offloading delay perceived by the UEs. Note that we assume that the MEC server is collocated with the gNB as outlined in, e.g., [19, 101]. Apart from the offloading delay, also an energy consumed at the UE for the offloading itself (in this thesis denoted as the offloading energy) can be considered in the offloading decision [99].

To handle the mobility of the users and enable seamless task offloading two options are possible, selection of communication path to deliver the offloaded task to the MEC and collect the results back, and allocation of computation resources (VMs or containers).

2.2.1 Communication with the MEC

Problem of path selection for scenario considering common mobile cloud computing is addressed in [102], where the authors propose to select the path using fuzzy logic. This idea covers selection of target cloud offloading system based on the path parameters such as delay, packet loss and benefits of offloading. However, this solution focuses only on centralized cloud services while radio aspects or mobility of users and possibility of handover are not reflected.

If we consider possible handover during transmission of data for computation, selection of the most appropriate way for data delivery to the computing BSs becomes problem analogical to routing in Wireless Sensor Networks (WSN). In this case, energy consumption on the side of SCeNB (containing the computation power for the MEC) is not such limiting factor as the SCeNBs are not powered by short life-time batteries. In WSN, plenty of algorithms have been defined. Basic routing algorithms for the WSN do not consider energy consumption of data delivery or dynamic path update [103]. In the MEC, the

energy is limiting only for radio communication between the UE and the SCeNBs. Also, dynamicity of the system is inherent feature of the mobile networks. Therefore, energy as well as dynamicity must be taken into account. The dynamicity of scenario for the WSN is addressed by Ad-hoc On-demand Multipath Distance Vector with Dynamic Path Update (AOMDV-DPU) [104]. Additionally to hop count metric, the algorithm selects paths based on Received Signal Strength Indicator (RSSI). However, even selection of paths with good RSSI to avoid weak radio links does not guarantee minimal delay. In addition, the AOMDV-DPU does not consider transmission energy, which is essential in our case. Similar weakness prevents implementation of Adaptive Multi-metric Ad-Hoc On-Demand Multipath Distance Vector (AM-AOMDV) [105] to the MEC since it routes data based on RSSI, latency and node occupancy. Moreover, backhaul from the serving BS to the operator's core network is typically wired. In addition, if the serving BS selection is based on RSSI, the same path to the core network would be selected all the time disregarding the selected SCeNBs for computation and backhaul status. Hence, the WSN-like approaches cannot be easily applied to our problem. Designed path selection algorithm should take into account the UE's limited energy resources, radio and backhaul conditions, and the UE's requirements on maximal possible delay for data delivery to guarantee the QoS. In order to combine transmission delay and energy, Power and Delay-aware Multi-path Routing Protocol (PDMPRP) is proposed in [106]. The PDMPRP chooses multi-paths in order to minimize energy consumption without increasing delay. With respect to [103–107] where whole network is wireless, backhaul from the serving BS to the operator's core network is typically wired. In addition, if the serving BS selection is based on RSSI, the same path to the core network would be selected all the time disregarding the SCeNBs selected for computation and backhaul status.

In the MEC, the application is offloaded from the UE to the MEC hosts (represented by the SCeNB) if it is profitable from energy consumption and/or delay perspective [21, 108]. After selection of the SCeNBs, which take care of computation, data must be delivered to these BSs. Typically, the SCeNBs are connected to network through a low quality backhaul comparing to common backhaul of the macrocell eNBs. Hence, distribution of data for computation from the BS providing radio access (denoted as serving BS) to all computing BSs through the backhaul of limited capacity (e.g., DSL) can lead to a significant delay. To that end, it is efficient to deliver data to the selected computing BSs not only through the serving BS but also via neighboring BSs provided that those are in the user's radio communication range. In the mobile networks, switching radio communication from the serving BS to another BS (labeled as target BS) in UE's neighborhood is known as handover. The purpose of handover in the mobile networks is to provide seamless connection to moving users. The handover is usually initiated according to radio channel quality offered by the serving and target BSs [109, 110], available capacity of backhaul [111], or energy consumption of the UE [112].

Authors in [113] propose three clustering strategies, which select a set of computing SCeNBs together with wired path (excluding radio) to computing cluster. The objective of these clustering strategies is to minimize either cluster latency, cluster power consumption,

or SCeNB power consumption. Contrary to [113], we focus on minimization of energy consumed at the UE and possibility to change radio path (between the UE and the SCeNB) for distribution of parallelized computation at several SCeNBs. In addition, our approach considers jointly energy consumed by the UE and transmission delay.

To satisfy even high demands of the devices on computation, the computing power distributed over nearby BSs. At each cluster, VM [13] or a container [14] can be run to host the UE's application. The VMs or container are deployed at the BSs with respect to their communication and computation capabilities. Selection of the MEC host forming the computing cluster and management of the computation according to the overall state of the network (i.e., current radio, backhaul and VMs state) is done by a Mobile Edge Orchestrator (MEO) [16, 19]. Apart from the MEO, an important element is a MEC platform manager, that is in charge of functionalities related to user service delivery, such as application running and scheduling. In existing approaches focusing on task offloading into the MEC, the data to the computing BSs is always delivered through the static serving BS [19, 21]. It means the UE is still attached to the same BS during delivery of whole offloaded data. Then, the serving BS distributes data through operator's core network to the computing BSs. This approach can be efficient if both radio channel between the UE and its serving BS as well as backhaul connection of the serving and all computing BSs are of sufficient throughput. Otherwise, a limitation at any part of the communication chain leads to a prolongation of the overall delay due to computation offloading.

2.2.2 Allocation of computing resources in MEC

The computing resources for the UE exploiting the MEC can be allocated at a BS that is the UE's serving BS [21]. However, as the UE moves, the distance and communication delay to the MEC host with the allocated computing resources increases. Thus, reduction of the communication delay can be done by migration of the computing resources closer to the UE. This approach exploits the ability to migrate the VMs [13] or the containers [14, 114] from one BS with MEC host to another one. However, this approach can pose a serious delay if not planned properly [95]. This is due to the delay of the VM migration [22], which can make a service unusable, if the VM migration is started when the UE's offloaded task is being processed. Another option of the mobility support is, instead of migrating the VMs to deploy an entirely new VM at the new BS and start the computing over [23]. This approach is denoted as a VM deployment. A similar approach is possible for the containers, which can exploit their advantage of a lower startup time compared to the VMs [95]. For the containers, the process is denoted as a container deployment [14]. The deployment of the new VM or container at another BS, however, leads to a wasting of energy and the computing resources of the MEC servers. Consequently, the performance of such approach becomes limited in scenarios with a heavy computation load. Therefore, with focus on offloading of real-time applications, the VM assigned to the UE should be ready when the computing task is being offloaded.

A straightforward approach, lies in migrating the VM to a new serving BS when the

UE changes its serving BS. In this case, the VM is migrated closer to the UE, leading to reduction in communication delay. In [115] the authors investigate whether it is efficient to migrate the VM from one BS to another during the movement of the UE.

The problem of joint computing and communication resource allocation for the MEC services can be solved by an iterative algorithms as proposed in [116, 117]. An extension of the iterative algorithm towards a distributed solution, which can be run on each BS separately is presented in [118]. In [118], the authors show that the performance of the distributed solution is close to the centralized one while collection of all information at the central control node in the network is not required. The computing and communication resource allocation with an interference management is proposed in [119]. The developed offloading decision is followed by a resource allocation with interference management exploiting graph coloring. In [120], the authors design two solutions for an optimal VM placement, based on the integer linear programming, to minimize the number of pre-allocated VMs and the degradation of the Quality of Experience (QoE). The energy consumption of the UEs is considered in [121] and [122]. The authors of [122] formulate the resource allocation problem as a convex optimization problem for a minimization of the UEs' energy consumption under a constraint on the computation latency and on the fairness of resource allocation. Then, an optimal policy for the resource allocation is derived. In [123], the authors propose a Q-learning-based algorithm for the resource allocation. The algorithm learns how to allocate the resources, and allocates these resource based on the actual state of the VM. Furthermore, in [124], the authors consider also a content caching for the resource allocation in order to improve the performance of the MEC. The caching is exploited to keep the content requested by the UE at the BS to alleviate the BS's backhaul. The authors formulate their optimization task as a convex problem, which is then transformed into a distributed convex optimization problem.

All the above-mentioned papers [116, 118–124] focus on the static UEs or the UEs with a very low mobility. However, a support for the mobility management of the UEs during the offloading is a key feature required to ensure seamless exploitation of the MEC services [125]. The solutions developed for the static UEs in [116, 118–122, 124] cannot be easily extended to support the UEs' mobility as the VM placement would have to be determined every time the UEs' positions change. Similar approach for determination of an optimal VM placement every time the UEs' positions change is considered in [126]. The authors exploit only the computing resources of the UEs. This is, however, not an easy task (if not infeasible) for the moving UEs due to the computation complexity of these algorithms and due to the fact that these papers do not consider an impact of the UEs' mobility on communication and computation.

To handle the UE's mobility in the MEC, dynamic algorithms are required. The dynamic algorithms based on the VM and considering the mobility of UEs are outlined, e.g., in [127] and [128]. The VM is migrated to a new BS whenever the UE changes its serving BS. This means that the VM is migrated to remain in a proximity of the UE in order to reduce the communication delay. The authors in [115] use MDP along with a threshold policy-based mechanism to optimize the VM migration. The proposed

algorithm is designed for 1-D mobility model without consideration of energy consumption and actual path selection. The algorithm from [115] is then extended to in [129], where the proposed solution decides whether and where to migrate the VM. The authors still consider Euclidean distance as the sole metric for decision on the VM migration. This work is further enhanced by a mobility prediction with a fixed accuracy [130] and consideration of the number of UEs utilizing the VM's resources at a given BS as a metric for decision on VM placement. Nevertheless, the algorithm for VM placement proposed in [130] delivers offloaded task via serving BS selected according to radio channels. This work is further enhanced in [130], where is assumed, and the computation load of the BS is considered as the decision metric on the top of the Euclidean distance. Another approach for the decision on the VM migration exploiting mobility prediction is presented in [131], where the authors propose a Q-learning-based algorithm determining the time when the VM migration should be started. The prediction of the UEs' mobility is critical for the VM migration, since the migration is both computation and communication resource demanding. Therefore, the VM migration-based solutions, exploited in [127–131] impose a significant delay (in order of seconds), which prevents their exploitation for the real-time applications [98].

Chapter 3

Thesis objectives

Based on the motivation, and state of the art, we specify objectives to satisfy the requirements on the future mobile networks. The objectives are defined as:

Objective 1 The mobile network optimization exploits mobile network and user information, that should be collected. Thus, a solution to enable the collection of the UEs' information from a large number of devices is defined as the first objective.

Objective 2 The collection of the UE' information can rely on the data relaying technique, but the UEs should be motivated to provide the energy of their UEs for the relaying purpose. Therefore, the objective is to design a cooperative communication that motivates the UEs to cooperate in relaying and to show that the relaying leads to a reduced energy consumption of the UEs.

Objective 3 The collected information is then exploited for an optimization of the mobile network. One of the challenges is to satisfy the UEs' data rate requirements. Therefore, the objective is to design a solution improving the UEs satisfaction with the QoS in terms of the data rate requirements.

Objective 4 In order to achieve real-time computation offloading for the UEs exploiting the MEC a low delay communication is necessary. Thus, the objective is to design the communication resource allocation algorithm that reduces the communication delay.

Objective 5 In the computation offloading, the MEC exploits not only communication resources, but also, computation resources. Thus, the last objective of this thesis is to propose a solution for joint allocation of both communication and computation resources to provide a seamless real-time offloading.

Chapter 4

Collecting user and network information

With the goal of self-optimized 3GPP mobile networks, a huge amount of data have to be collected and processed. To enable collection of such a huge amount of data, several changes to the current 4G mobile networks have to be done. Therefore, in this chapter, we describe the changes necessary to enable communication of a large number of devices to collect mobile network information from the UEs for the network optimization. The content of this chapter is based on [132–135].

The communication in the 4G mobile networks is in general done via the UE transmitting its data to its serving BS, and receiving data directed to the UE from the BS. However, this is not always beneficial, when the amount of data transmitted from the UE to its serving BS is small. Therefore, a different approach that provides a suitable option for collecting data from the UEs is proposed in the Section 4.1. To show the benefits of the proposed solution is compared to the existing approaches in performance evaluation. The solution provided in the Section 4.1 exploits D2D, therefore, in the Section 4.2 a resource allocation based on the NBS is proposed to motivate the UEs to cooperate. Then, in the Section 4.3 energy consumption analysis of the relayed communication via the D2D is provided to show its benefits. In the Section 4.4 we propose algorithms for joint positioning of the FlyBSs and UEs association that exploits the collected mobile network information.

4.1 Cross-layer optimization of LTE-A signaling

One of the major obstacles in the collecting information from the UEs is the communication overhead, which impact increases with decreased size of the transmitted payload, i.e., useful data. Thus, first objective is to reduce overhead and keep its amount constant for each transmission in order to simplify scheduling of the communication resources. Therefore, we target to use semi-persistent scheduling (i.e. each device communicates over the same RBs periodically, but some changes can be made) without the need for

additional resource allocation using non-persistent scheduling (each time device communicates, resources have to be allocated again). To enable semi-persistent scheduling, we propose new signaling (described latter) that enables the device to inform the eNB about the payload size in more precise way than the BSR [41]. Overhead is further reduced by using buffering and enabling collection of payload from more devices. Buffering allows us to send multiple payloads per single signaling message. However, as mentioned before; we have to respect the Time To Live (TTL) of the payload [136].

To send more payloads at once, clustering concept is exploited. The clustering enables to form clusters of nearby devices via D2D relaying communication. For each cluster, a cluster head is selected out of all devices in the cluster. The cluster head collects payloads from the devices within the cluster and transmits them to the eNB. The clustering is further enhanced by buffering as shown in Figure 4.1. The cluster head (in Figure 4.1 denoted as *cl_head*) buffers payloads from the devices within the cluster. The devices inform the cluster head about TTL of their payloads in order to schedule transmission of individual payloads properly. This information is delivered from the device to the cluster head by means of D2D communication. To transmit buffered payloads to the eNB, signaling message is added and sent by the cluster head. The scheme merging new signaling, buffering, and clustering is labeled as Cross-layer Optimization (CLO).

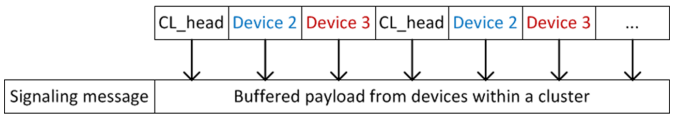


Figure 4.1. Principle of buffering within cluster.

Both clustering and buffering reduce the number of transmissions from devices to the eNB. Clustering reduces the number of transmitting devices in the space domain (devices within a specific area transmit as one device to the eNB) while buffering in the time domain (device transmits once per multiple TTL). Therefore, they can be considered complimentary.

In Figure 4.2, we show high-level overview of the proposed approach for collection of payloads from devices in the network. We assume a single eNB to which all devices are connected, either directly or via the cluster head. In this proposal, we assume basic clustering to show lower-bound of the gain introduced by our proposed scheme. Clustering is, therefore, based on distance (cl_{dist}). It means that the cluster is formed as a set of devices with mutual distance up to cl_{dist}. Advanced clustering approach can further improve performance, but it is left for future research. In Figure 4.2, DEV denotes device and represents common user's device, such as smartphone or tablet, as well as a sensor or a machine. The DEV_4 and DEV_8 are the cluster heads of Clusters 1 and 2, respectively, as they are closest to the eNB. The DEV_5 and DEV_6 are not members of any cluster as they are not in vicinity of other devices. These devices can be seen as the cluster heads of their own clusters with only themselves in the cluster. If a new DEV would require transmission of its payload within the proximity of DEV_5 or DEV_6, it could join either DEV_5 or DEV_6 and form a new cluster together.

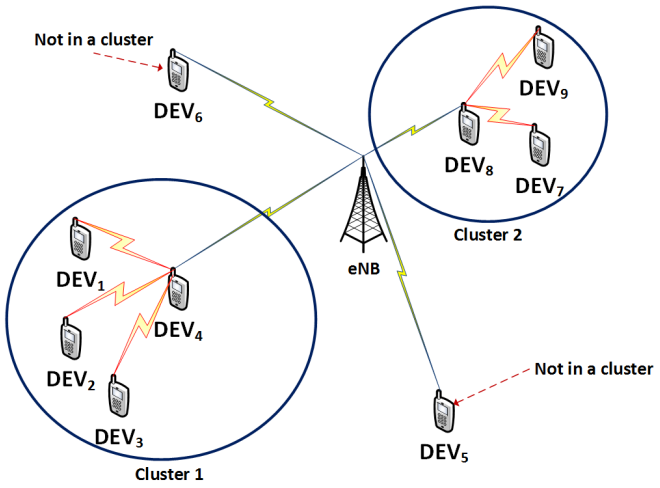


Figure 4.2. Scenario of the proposed approach for collection of information from devices.

4.1.1 Management of the proposed scheme

In this subsection, we describe management procedure of the CLO as shown in Figure 4.3. This figure shows a procedure, followed by each device, which wants to start sending its payloads. The procedure begins with checking whether there is a cluster head in the proximity of the device willing to transmit data. Based on this checking, there are three options for the device: a) the device becomes the cluster head if there is no cluster head in the vicinity, b) the device joins existing cluster, or c) the device is selected as the new cluster head for existing cluster. Communication between devices within the cluster exploits D2D communication. Using D2D communication, the devices form and manage clusters and transmit their payloads to the cluster head [137]. In case (a), where no cluster exists in the device's vicinity; the device buffers its payloads (respecting TTL of the content) and then starts RAP to obtain radio resources. Afterwards, the control message and payload are sent. Finally, semi-persistent scheduling is initiated and the device sends buffered payloads. In case (b), the device joins existing cluster head and starts transmission of the payloads to the cluster head using D2D communication. In the case (c), when the device is selected as the new cluster head (the device is closer to the eNB than the existing cluster head), all devices within the cluster are informed about new cluster head. This information is issued by the former cluster head. After this, the device, which becomes the cluster head starts receiving payloads from the devices within the cluster and initiates RAP. Then, the cluster head sends control message and initiates semi–persistent scheduling. Finally, the cluster head transmits collected payloads to the eNB.

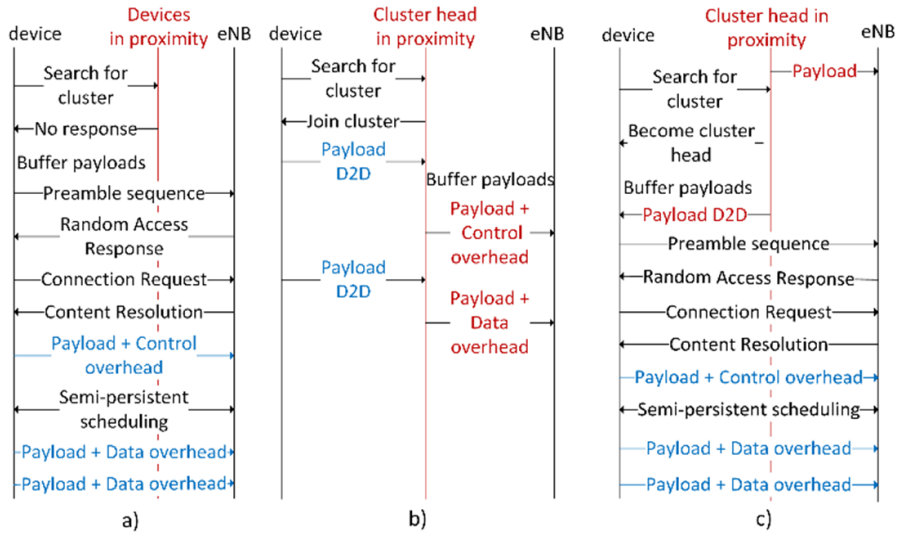


Figure 4.3. Procedure for the device beginning its transmission of payload in case of (a) no cluster in proximity, (b) joining cluster in proximity, (c) becoming cluster head.

The CLO defines four types of signaling messages in order to replace ROHC. Two types of control messages are intended for initial transmission of the stand-alone device (option a in Figure 4.4) and cluster head (option b). The other two options (c and d in Figure 4.4) are designed for transmission of the data from the stand-alone device (option c) and cluster head (option d). Using the control messages, the eNB initiates a record for the further transmissions of the device. This record is stored in a database at the eNB to reconstruct the full TCP/IP header if the device's payload designation is in the Internet. The control message is always send as the first message. Then, data message is send in the subsequent transmissions. The first field in the proposed signaling messages is D/C. It specifies the type of the message in order to distinguish between data and control messages. The second field, U/CL, denotes whether the device transmits data to the eNB by itself or via the cluster head. These first two fields are the same for all four types of messages. Following fields in the signaling messages are defined depending on the type of message. Fields $DEST$ and SRC denote destination and source address for the payload, SEQ is the sequence number of the transmitted payload. Information about the payload size is carried in $Payloadsize$ field. The last field of each message is Cyclic Redundancy Check (CRC), which ensures correct delivery of the signaling message. For communication within the cluster, i identifies each device and ranges from 1 to the number of devices within the cluster (N_DEV). The flag A/S is used to inform the eNB, that payload of each device within the cluster is included, or if the payload from selected devices is included. If payloads from not all devices are included, field $Bitmap$ is included. This field identifies devices, from which the payloads are being transmitted.

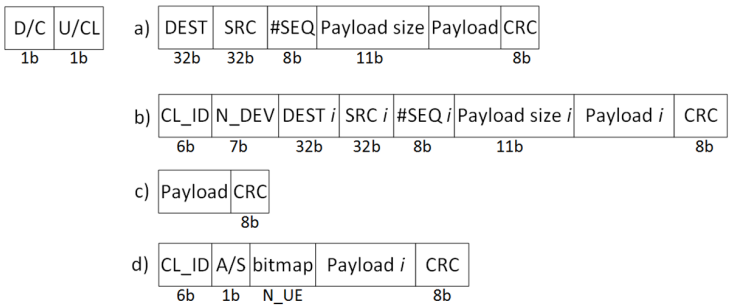


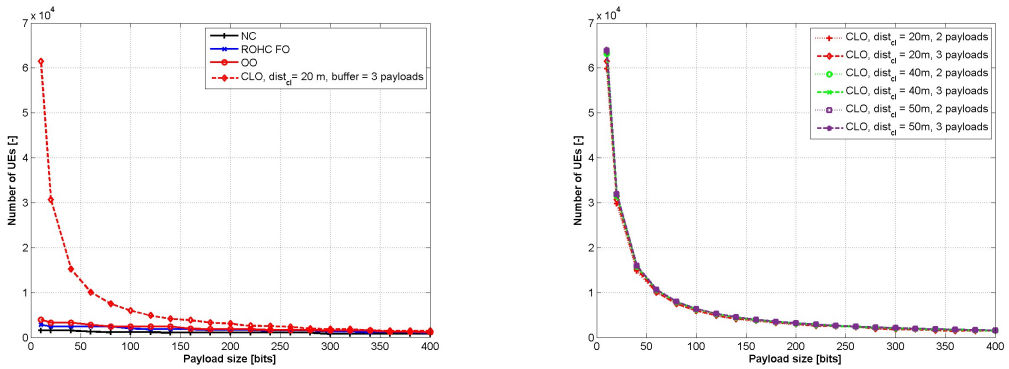
Figure 4.4. Proposed signaling messages enabling cross-layer optimization of frequent transmission of small payloads.

To assure correct order of transmitted packets and reception of every transmitted packet if this is required by the service or application, we utilize Hybrid Automatic Repeat reQuest (HARQ) and Automatic Repeat reQuest (ARQ) to check and repair received data. This enables to send constant message size after the control transmission and simplifies the semi-persistent scheduling. Thus, we determine resource allocation once and we do not use non-persistent scheduling even if reliable delivery of data is required as the ROHC does.

4.1.2 Evaluation of the proposal for collection of data from devices

In this subsection, we analyze the number of devices that can be served in mobile networks and ratio of the signaling overhead. We compare the proposed CLO with two schemes: i) scheme without any overhead compression, i.e., sending full TCP/IP overhead with a size of 40 bytes (labeled as No Compression (NC) in following figures); and ii) the ROHC in the ROHC First Order (ROHC FO) state [43]. The ROHC FO is selected instead of the ROHC Second Order (ROHC SO) as ROHC FO is send 5 times per every 100 packets [43]. The ROHC in the Initialization and Refresh (IR) state is not shown as its signaling is larger than for the NC [43]. We further include also results for our proposed signaling replacing of the ROHC but without buffering and clustering. This scheme is denoted as Overhead Reduction (OR) in all following figures. The LTE-M is not considered for performance comparison as it adopts different multiplexing and it is not backward compatible with the 4G. Note that the overhead in our simulations contains overhead introduced by all layers (TCP/IP, Packet Data Convergence Protocol (PDCP), MAC, and Radio Link Control (RLC)).

In Figure 4.5a, we show how many devices can be served by one eNB. With increasing payload size, the number of served devices decreases because more resources are required for the transmission of all devices. The NC enables eNB to serve the lowest amount of devices (1537 devices for 10 bits payloads) comparing to other schemes. The ROHC FO roughly doubles the number of served devices against the NC (gain up to 110.5%). The proposed OR, which is based only on optimization of overhead of ROHC without considering clustering and buffering, improves the number of served devices by additional up to

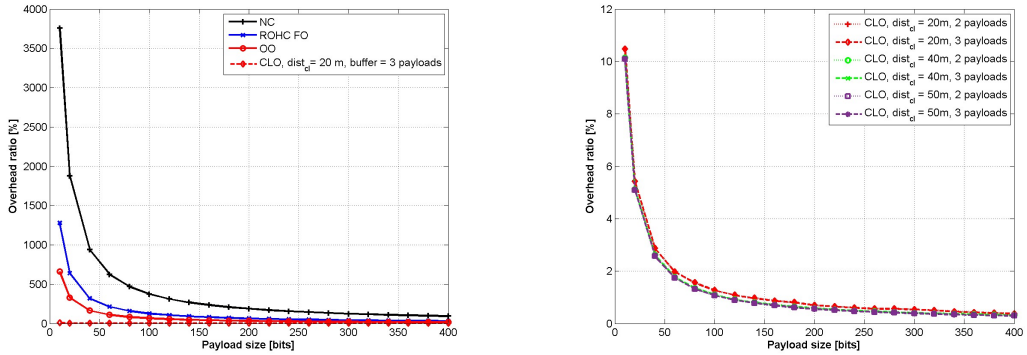


(a) (b)

Figure 4.5. The number of served devices transmitting frequently small payloads (a) comparison of the proposal and competitive schemes, (b) impact of clustering and buffering.

40% comparing to the ROHC FO. Far the best performance is achieved by the proposed CLO. The gain is more significant for small payloads as the payloads from more devices can be buffered together and sent within one transmission. The CLO enables to serve more than 65 000 devices transmitting 10 bits payloads. It corresponds to improvement in the number of served devices up to 22.8 times and 16.3 times compared to the ROHC FO and the OR, respectively. This gain is a result of sending more payloads from nearby devices in one message, which is achieved by the combination of clustering and buffering. Moreover, by clustering, only the devices closest to the eNB (using higher MCS) transmit and, thus, less resources are required for the transmission. From the results, we see that existing solutions NC and ROHC FO are not suitable for the 5G as the number of devices served if these approaches are adopted is lower than the expected number of devices connected to one eNB in 5G (10 000 to 100 000 devices, see [2]). However, our solution with only basic, not optimized, clustering enables to serve the required number of devices even for 10 MHz bandwidth, which is much lower that the bandwidth expected for 5G. Further increase in the number of served devices by our proposed approach can be reached by simple extension of bandwidth. This also shows that we can serve the required amount of devices with lower density of eNBs. Hence, the overall cost of the network deployment required for IoT or MTC can be lowered.

In Figure 4.5b, we show the impact of the number of buffered payloads and cluster radius on the number of the served devices. Increase in the payload size leads to decrease in the number of served devices as more resources are required for the transmission. Impact of increasing number of buffered payloads and increasing cluster radius on the number of served devices is negligible as the difference is less than 7.2%. This 7.2% improvement represents further increase in gain with respect to ROHC FO in the number of served devices so that the proposed CLO increases the number of served devices by up to 24.4 times for the most efficient buffering and clustering combination (cl_{dist}=50m, 3 payloads). Impact of clustering is limited by cluster size (number of devices within the cluster) as a large cluster leads to the same problem as large number of devices. Impact of buffering



(a) (b)

Figure 4.6. Overhead ratio (a) by the proposal and competitive schemes, (b) impact of clustering and buffering.

is, on the other hand, limited by a need to respect TTL of the transmitted data.

In Figure 4.6a, we compare overhead ratio, i.e., the ratio between the overhead and the payload. As expected, the overhead ratio decreases with increasing payload size. The ROHC FO decreases overhead ratio by 66% comparing to the NC. Further decrease in the overhead ratio is introduced by the OR. The OR reduces the overhead ratio by 48.5% comparing to the ROHC FO. However, still, the OR leads to significant ratio of the overhead to the payload (660% for 10 bits payload). Significant improvement is reached by the CLO, which reduces the overhead ratio to less than 10.5%. It corresponds to up to 68 times reduction comparing to the OR, up to 132 times comparing to the ROHC FO and up to 390 times comparing to the NC.

In Figure 4.6b, we show the impact of parameters of the CLO (number of buffered payloads and cluster size) on the overhead ratio. Difference in the overhead ratios between configurations of parameters is minimal like for the number of devices. Improvement by using various cluster sizes or numbers of buffered payloads is less than 0.3

In Figure 4.7, we show the impact of different duty cycle time (interval between two consequent payloads generated by one device) on the number of the served devices for payload of 100 bits. The number of served devices increases linearly with duty cycle as the devices generate payload less often.

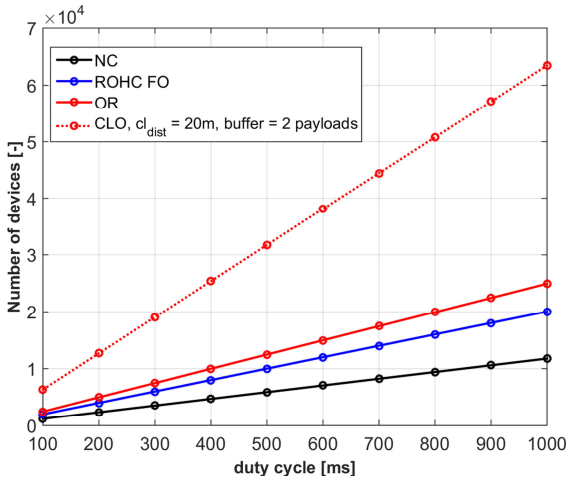


Figure 4.7. Impact of clustering and buffering on overhead ratio.

More details about the proposed solution can be found in [132].

4.1.3 Conclusion

In this chapter, we have proposed a cross-layer solution to increase the number of devices that can be served by one eNB. The solution combines reduction of the TCP/IP overhead with buffering and clustering concepts in order to maximize efficiency of the transmission of small payloads by a high number of devices such as sensors, machines, or conventional user devices. The proposal enables to serve more than up to 65 000 devices by one eNB in case of a 10 MHz bandwidth. This represents 24.4 times increased number of devices with respect to the state of the art solutions. Even if the proposed solution is compatible with existing 4G networks, it enables to serve the number of devices expected to be connected in 5G networks only with 10 MHz bandwidth. Therefore, the proposed solution enables collection of the information required for real-time optimization of the mobile networks.

4.2 Cooperative resource allocation in a relayed communication

In the previous section, we have proposed a solution to increasing number of devices (UEs, vehicles, sensors, etc.) that exploits transmission relaying via the D2D communication. However, the device that provides communication relaying should be motivated, as the relaying consumes additional energy, on top of the energy consumed by its own transmissions. Therefore, in this section, we derive a NBS for allocation of communication resources such that all devices in the network benefit from the relaying. Unlike related works, our NBS is based on energy consumption of the communication and is solved for N devices. Moreover, we solve the NBS for a general relaying strategy, and thus, the described solution is independent of actual relaying strategies considered by the devices.

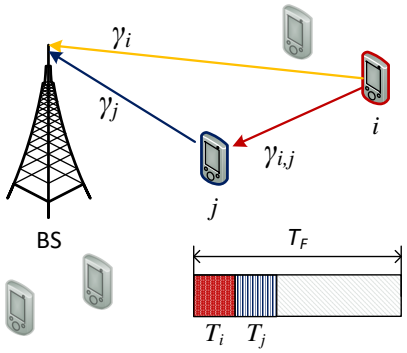


Figure 4.8. System model with device i relaying its data through device j.

In contrast to [9] the proposed solution is in closed form and does not require an iterative approach or auctions to reach the optimal solution. Thus, the NBS is applicable to wireless communications even with rapidly time-varying radio channels and high number of devices due to low complexity and high scalability.

4.2.1 System Model

Consider a single BS, which serves N devices (mobile phones, sensors, etc.). Each device transmits packets with M bits of data to the BS in the uplink direction. We focus on a case where the devices can act as relays through which other devices transmit their data to the BS as shown in Figure 4.8. The relays exploit Decode and Forward (DF) relaying scheme. Our system model is based on the system model exploited for example in [138].

In the considered scenario the devices share radio resources by means of Time Domain Multiple Access (TDMA). The devices compete for a part of a frame with a duration of T_F. Each device transmits for a portion of T_F defined as TTI $T_i = \alpha_i T_F$, where $\alpha_i \in (0, 1)$ and $\sum_{i=1}^{N} \alpha_i = 1$. Note that the proposed solution for TDMA can be extended towards OFDMA, but we leave this extension for the future due to limited space.

Communication between a source device (i.e., the device, which is willing to transmit the data) and the BS is done either by a direct communication or by a relaying via another device. For the direct communication, the data is transmitted by the i-th device (source) to the BS and the BS receives data with SINR γ_i. In case of the relaying, the i-th device (source) transmits data to a selected j-th device (relay) over a D2D channel. The j-th device receives the data with SINR $\gamma_{i,j}$. Then, the relay forwards the data of the source device to the BS over its direct channel and SINR at the BS is γ_j.

The data rate of the i-th device communicating directly to the BS is $r_i^d = B\log_2(1+\gamma_i)$, where B is the bandwidth allocated for the direct communication with the BS. The data rate between the source and the relaying devices is defined as $r_{i,j}^{D2D} = B^{D2D}\log_2(1+\gamma_{i,j})$. The data rate $r_{i,j}^{D2D}$ of the i-th device to the j-th device can be higher than the data rate achievable by the j-th relay at its direct channel to the BS (r_j^d). Thus, we adapt the data

rate at the relay channel to match data rate at the direct channel of the relay, i.e., the data rate at relay channel is $\bar{r}_{i,j}^{D2D} = min(r_j^d, r_{i,j}^{D2D})$.

The energy consumed by the direct transmission of a packet is expressed as:

$$E_i^d = \frac{(P_i^{\text{tx}} + P_i^c)\, M_i}{r_i^d} \tag{4.1}$$

where P_i^{tx} is the power consumed for the transmission, P_i^c is the power consumed by the circuitry of the i-th device, and M_i is the amount of bits to be transmitted by the i-th device.

The energy consumed by the D2D transmission of the packet from the i-th device (source) to the j-th device (relay) is then expressed in similar way, i.e.:

$$E_{i,j}^{D2D} = \frac{\left(P_i^{\text{tx,D2D}} + P_i^c\right) M_i}{\bar{r}_{i,j}^{D2D}} \tag{4.2}$$

where $P_i^{\text{tx,D2D}}$ is the power consumed by the D2D transmission of the i-th device. The i-th device can transmit its data directly or via j-th relay by exploiting the relaying strategy s_i from a set of possible strategies $\mathbf{s}$, given as:

$$s_i = \begin{cases} j & \text{if transmitting via } j\text{-th device} \\ 0 & \text{otherwise (direct trasnmission to the BS)} \end{cases} \tag{4.3}$$

If the device decides not to follow the relaying strategy (i.e., $s_i = j$), it follows a disagreement strategy d (i.e., $s_i = 0$). Under the disagreement strategy d, the device does not cooperate with others and transmits data directly to the BS, disregarding strategies of other devices.

Based on the strategy selected by the device, we define the energy consumed for transmission of the i-th device following the strategy s_i as:

$$E_i^{\text{tx}}(s_i) = \begin{cases} E_{i,j}^{D2D} & \text{if } s_i \neq 0 \\ E_i^d & \text{otherwise} \end{cases} \tag{4.4}$$

Note that the energy consumed by the relaying devices for reception is omitted in the model as it leads to different solution, which is more complex and does not fit to the page limit. We assume that each device has initial energy E_i^{init}. Then, the total number of packets transmitted by the device following s_i before the battery depletion is defined as:

$$N_i(s_i) = \frac{E_i^{\text{init}}}{E_i^{\text{tx}}(s_i)} \tag{4.5}$$

The coordination of resource allocation is done in a central way by a BS, as described in, e.g., [139]. The only information needed to be collected by the BS is either energy consumption or channel quality, which is anyway reported to control the communication

in the mobile networks even with D2D relaying.

4.2.2 Problem Formulation

Our objective is to allocate TTIs to the devices under cooperation via Nash Bargaining solution. The Nash Bargaining solution is a class of a cooperative games where each player follows strategy, which reaches a mutual agreement among the players and has a higher utility than a non-cooperative strategy.

Let $\mathbf{N} = \{1, 2, \ldots, N\}$ be a set of players, in our case represented by the devices willing to transmit data. Let $\mathbf{Q}$ be a closed and convex subset of $\mathbb{R}^N$ representing the set of feasible payoff allocations that the players can get by cooperation. Then, let $\mathbf{A} = \{\alpha_1, \alpha_2, \ldots, \alpha_N\}$ be a set of feasible allocations of TTIs to the devices. Let $\bar{N}_i(d) = \frac{N_i(d)}{N}$ be the minimal payoff required by the i-th player, otherwise, the i-th player does not cooperate. Suppose $\left\{\alpha_i N_i(s_i) \in \mathbf{Q} | \alpha_i N_i(s_i) \geq \bar{N}_i(d), \forall i \in \mathbf{N}\right\}$ is a nonempty bounded set. We define $\bar{N}(\mathbf{d}) = \left(\bar{N}_1(d), \ldots, \bar{N}_N(d)\right)$, then the pair $\left(\mathbf{Q}, \bar{N}(\mathbf{d})\right)$ is called the N-person bargaining problem.

The objective is to find the NBS of the TTIs allocation $\mathbf{A}^*$, which maximizes the product (benefit) of the number of transmitted packets gained by the cooperation. This objective is formulated as:

$$\mathbf{A}^\star = \arg\max_{\mathbf{A}} \prod_{i=1}^{N} \left(\alpha_i N_i(s_i) - \bar{N}_i(d)\right) \tag{4.6}$$

$$\text{subject to } \alpha_i^\star N_i(s_i) \geq \alpha_i N_i(s_i), \forall i \in \mathbf{N} \tag{4.7}$$

$$0 < \alpha_i < 1 \tag{4.8}$$

$$\sum_{i=1}^{N} \alpha_i = 1 \tag{4.9}$$

The constraint (4.7) motivates devices to cooperation as it specifies that the number of transmitted packets for each device following α_i^* must be higher than if the device would follow any other α_i. The constraint (4.8) limits α_i to allocate each device a portion of TTIs while the constraint (4.9) guarantees that the resources allocated to all devices fit to a single frame.

4.2.3 Nash Bargaining Solution

In this section, we first derive the NBS for two devices. Then, we generalize the solution towards N devices. Since the objective function (4.6) is convex, we explore the Karush-Kuhn-Tucker (KKT) conditions for two as well as for N devices.

NBS for two devices

In this subsection we consider two devices i.e., $N = 2$, in line with model in Figure 4.8. In this case, a device which provides relaying, i.e., Device 2 (j), is allocated with a fraction

α_2 of the frame and a device exploiting relaying, i.e., Device 1 (i) gets $\alpha_1 = 1 - \alpha_2$ of the frame by Pareto optimality. To derive the NBS, we formulate the Nash product in terms of α_2 (where the constraints are already incorporated by the choice of α_1 and α_2:

$$L = \left[(1-\alpha_2)N_1(s_1) - \frac{N_1(d)}{2}\right]\left(\alpha_2 - \frac{1}{2}\right)N_2(s_2) \tag{4.10}$$

Then, the derivative of L with respect to α_2 is set equal to zero:

$$\frac{\partial L}{\partial \alpha_2} = \frac{N_2(s_2)\left[N_1(s_1)\left(3-4\alpha_2\right) - N_1(d)\right]}{2} = 0 \tag{4.11}$$

By solving the linear equation in (4.11) for α_2 and substituting $\alpha_1 = 1 - \alpha_2$ we obtain the NBS for TTI allocation for both devices where

$$\alpha_1 = \frac{1}{4} + \frac{E_2^{\text{tx}}(s_2)}{4E_2^{\text{tx}}(d)} \qquad\qquad \alpha_2 = \frac{3}{4} - \frac{N_2(d)}{4N_2(s_2)} \tag{4.12}$$

The numerical analysis of the number of transmitted packets is done in a scenario with parameters from [45], i.e., $P_i^{tx} = P_i^{\text{tx,D2D}} = 200\,\text{mW}$, $P_i^c = 800\,\text{mW}$, $B = B^{D2D} = 200\,\text{kHz}$, $M = 100\,\text{B}$, $E_i^{init} = 100\,\text{J}$, and $T_F = 10\,\text{ms}$. The derived NBS is compared with *Equal* TTI allocation when each device is allocated with $\frac{1}{N}$ of T_F, *MaxMin* TTI allocation, where the minimal number of transmitted packets per a device is maximized [54], and to *Direct* transmission scheme without relaying, where each device is allocated with $\frac{1}{N}$ of T_F. The derived NBS works independently of relaying strategy. Nevertheless, for comparison with other allocations, we select a commonly exploited Opportunistic Relay Selection (ORS) [47]. This strategy considers quality of both the direct channel (γ_i) and the D2D channel between source and relay devices ($\gamma_{i,j}$) for selection of the relaying device, i.e., the strategy s_i for the OR is defined as:

$$s_i = \arg\max_{j \in N} \min\left(\gamma_j, \gamma_{i,j}\right) \tag{4.13}$$

In Figure 4.9, the number of transmitted packets is shown for the Device 1 (N_1) and Device 2 (N_2) as a function of $\gamma_{1,2}$. The Device 2 acts as the relay for the Device 1. Note that the packets from Device 1 relayed by Device 2 are not included in the number of packets transmitted by the Device 2 (i.e., in N_2). For all three relaying algorithms, N_1 increases with $\gamma_{1,2}$ due to improvement in the relaying channel quality. The MaxMin algorithm results in the highest N_1, but the lowest N_2 out of all relaying algorithms, because the MaxMin targets to provide fairness among the devices ($N_1 = N_2$ and lines for the Device 1 and the Device 2 overlap in Figure 4.9). As a result of fairness, the Device 2 does not cooperate since it looses with respect to the direct transmission. The Equal algorithm improves N_i for the Device 1 with respect to the direct transmission, however, the performance of the relaying Device 2 is the same as for the direct transmission. This means the Device 2 is still not motivated to cooperate and help the Device 1. In contract to this, the derived NBS results in a gain for both devices with respect to the

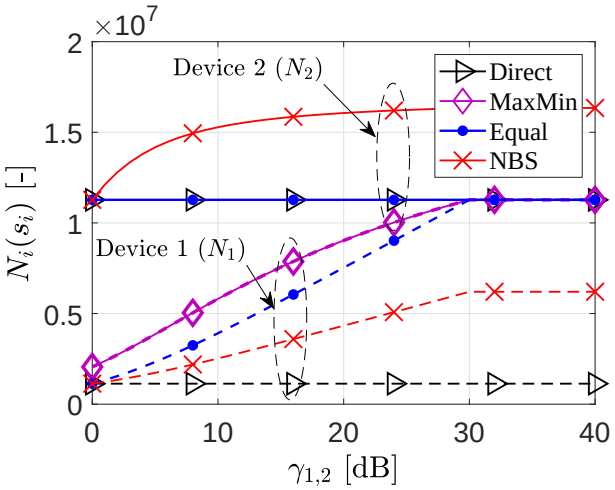


Figure 4.9. Number of transmitted packets ($N_i(s_i)$) with $\gamma_1 = 0$ dB and $\gamma_2 = 30$ dB, (dashed line - Device 1 (source), solid line - Device 2 (relay).

direct communication. Consequently, the Device 2 is motivated to cooperate with the Device 1, because the Device 2 receives an incentive in terms of additional resources for communication as a reward for its cooperation. Assuming rationality of players (devices), only the derived NBS leads to natural cooperation of devices.

NBS for N devices

In this subsection, we generalize the solution obtained for two devices towards N devices. First, we replace product in (4.6) by the sum of logarithms:

$$\mathbf{A}^\star = \arg\max \sum_{i=1}^{N} \log\left(\alpha_i N_i(s_i) - \bar{N}_i(d)\right) \tag{4.14}$$

Then the Lagrangian of (4.14) is derived considering conditions (4.7), (4.8), and (4.9):

$$L = \sum_{i=1}^{N} \log\left(\alpha_i N_i(s_i) - \frac{N_i(d)}{N}\right) + \mu\left(\sum_{i=1}^{N} \alpha_i - 1\right) \tag{4.15}$$

Taking the derivatives of the Lagrangian with respect to α_i and μ, and by setting the derivative equal to zero, we get:

$$\frac{\partial L}{\partial \alpha_i} = \frac{N_i(s_i)}{\alpha_i N_i(s_i) - \frac{N_i(d)}{N}} + \mu N = 0 \tag{4.16}$$

$$\frac{\partial L}{\partial \mu} = \sum_{i=1}^{N} \alpha_i - 1 = 0 \tag{4.17}$$

From (4.16) we obtain

$$\alpha_i = \frac{N_i(d)}{N_i(s_i)N} - \frac{1}{N\mu} \tag{4.18}$$

where μ is determined from (4.17) and (4.18) as:

$$\mu = \frac{1}{1 - \sum_{i=1}^{N} \frac{N_i(d)}{N_i(s_i)N}} \tag{4.19}$$

Next, by substituting (4.19) to (4.18), we obtain the NBS of TTI allocation in a closed form as:

$$\alpha_i = \frac{N_i(d)}{N_i(s_i)N} + \frac{1 - \sum_{i=1}^{N} \frac{N_i(d)}{N_i(s_i)N}}{N} \tag{4.20}$$

To obtain the allocations of TTIs to the devices in the terms of the transmission energies, we substitute the number of transmitted packets from (4.5) into (4.20):

$$\alpha_i = \frac{E_i^{tx}(s_i)}{E_i^{tx}(d)N} - \frac{\sum_{i=1}^{N} \frac{E_i^{tx}(s_i)}{E_i^{tx}(d)N} - 1}{N} \tag{4.21}$$

The derived allocation (4.21) is in closed form, which makes it suitable for wireless communications with a frequently varying quality radio channel. The complexity of (4.21) is $\mathcal{O}(N)$, thus the solution is suitable even for scenarios with a high number of devices, as envisioned in 5G mobile networks).

4.2.4 Simulation Results and Analysis

In this section, we present numerical results obtained by simulations following parameters defined in Section 4.2.3 in line with [45]. The results for the NBS are compared with all three allocation schemes (MinMax, Equal, Direct) described also in Section 4.2.3. The devices are uniformly distributed in a simulation area with diameter of 500 m around a single BS. The direct channel is modeled as Urban Macro with Log-normal shadowing with variance of 4 dB and the D2D channel follows Winner II model. Each device has the initial energy generated from exponential distribution with $\lambda = 1$ and maximal value of 100 J.

The average energy consumed per transmission $E_i^{tx}(s_i)$ is shown in Figure 4.10a. This figure, shows that energy efficiency is improved via relaying with respect to the direct transmission, disregarding whether the cooperation is natural (for the NBS) or must be externally enforced (for MaxMin and Equal).

Figure 4.10b shows the total number of transmitted packets over the number of devices deployed in the area, i.e., $\sum N_i(s_i)$. For the Direct transmission, the number of transmitted packets is almost constant disregarding the number of devices, because each device transmits data directly to the BS. For the MaxMin allocation, the total number of transmitted packets decreases with an increasing number of devices, as the MaxMin targets a fairness in N_i. For the Equal allocation and for the NBS, the total number of transmitted packets is increasing with the number of devices, because a higher number of possible relays can appear in proximity of the source device due to a higher number of devices in the area.

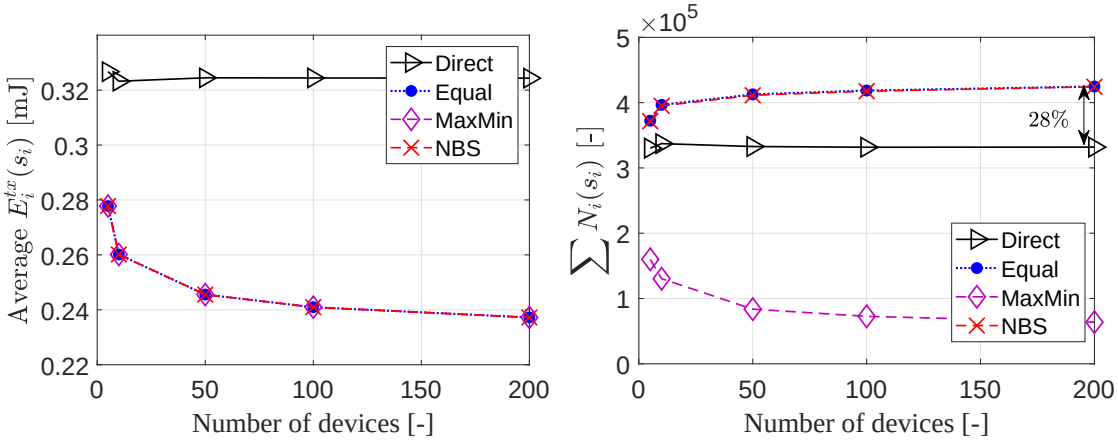


Figure 4.10. Average energy consumed per transmission ($E_i^{tx}(s_i)$) (a) and total number of transmitted packets ($\sum N_i$) (b).

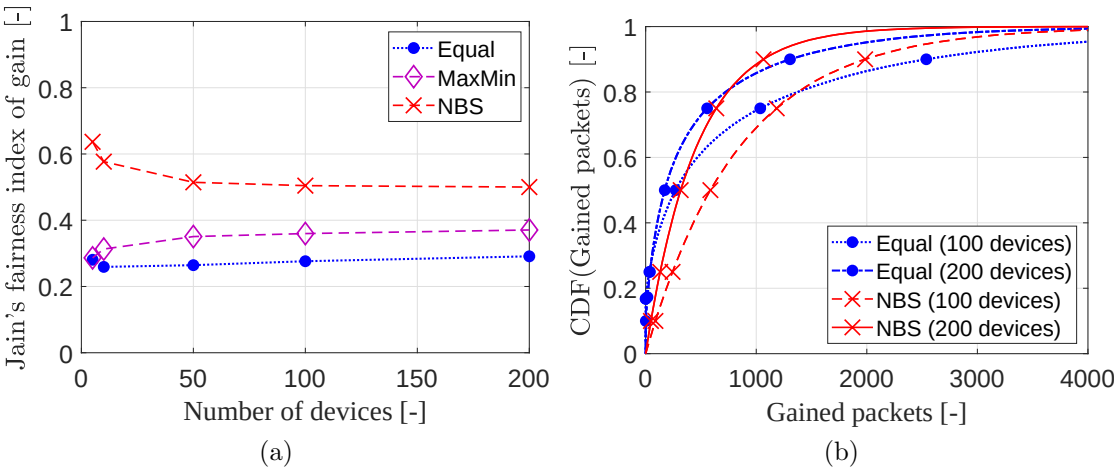


Figure 4.11. Jain's fairness index of number of transmitted packets gained by cooperation (a) and CDF of the gain (b).

In the Figure 4.11a, we show fairness in gained number of transmitted packets via Jain's fairness index. The fairness in gain is highest for the NBS, as the NBS motivates devices to cooperate via fair sharing of benefits by all devices. A lower fairness in the distribution of the gain among the devices for the MaxMin is a result of the fact that the algorithm targets fairness in N_i, but disregards gain in the number of transmissions by individual devices. The Equal allocation splits the time fairly, but disregards channel quality and provides the worst fairness in gain out of all the compared schemes.

Figure 4.11b shows that the NBS distributes the gain in the number of transmitted packets more fairly among the devices comparing to the Equal allocation for 100 and 200 devices.

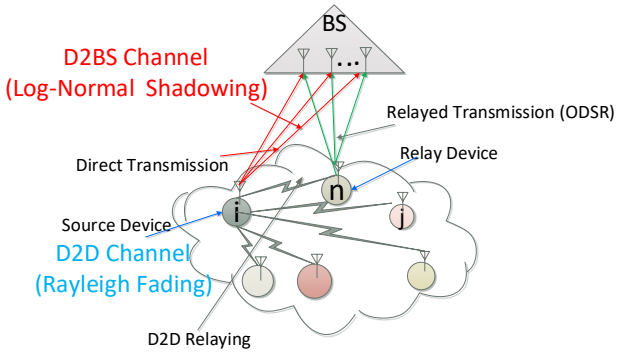


Figure 4.12. D2D relaying in the uplink communication of a single cell network. Devices are inside a shopping mall/university building/ offices and the BS is far away separated by walls. The devices have single antenna while the BS has multiple antennas.

4.2.5 Conclusion

In this chapter, we have derived energy consumption-based Nash Bargaining solution for allocation of communication resources maximizing the number of transmitted packets. The derived NBS motivates the devices to cooperation and provides a solution to the device clustering bases solution, proposed in Section 4.1. The NBS is in closed form, thus, it is applicable even to wireless communication with frequently varying channel. Due to a very low complexity, the derived NBS is scalable and suitable for scenarios with very high number of devices, as envisioned in 5G mobile networks. Furthermore, the derived NBS works independently on the relay selection algorithm.

4.3 Energy Consumption of Opportunistic D2D Relaying Under Lognormal Shadowing

The proposed clustering based solution with allocation of the communication resources via the derived NBS provide a suitable option for increasing number of communicating devices while considering the communication energy. However, the relay selection algorithm that not only increases amount of the collected information for the mobile network, but also maximizes the battery life time is necessary. Therefore, in this section, we provide a mathematical analysis of a distributed opportunistic relay selection to show its benefits for the collection of the mobile network information from the UEs.

4.3.1 System Model

For the analysis, we consider a single-cell network with a BS (equipped with $M \geq 1$ antennas) and N single-antenna devices for uplink data transmissions. The devices are uniformly distributed in the network. We focus on a two-hop transmission model, where a source device can either transmit data directly to the BS or relay the data to a nearby device, which forwards the data to the BS, as depicted in Figure 4.12.

In a direct transmission, the received signal vector at the BS from the i-th device is given as:

$$\mathbf{y}^{\mathrm{BS}} = \sqrt{P}\mathbf{h}x_i + w \tag{4.22}$$

where $\mathbf{y}^{\mathrm{BS}} = \{y_1, y_2, \cdots y_M\}^T$ is the $M \times 1$ received signal vector, P is the transmit power, x_i is the transmitted signal with unit power $\mathbb{E}[|x_i|^2] = 1$, $w \sim \mathcal{CN}(0, N_0)$ is the zero-mean Additive White Gaussian Noise (AWGN) with variance N_0, and $\mathbf{h} = \{h_{1i}, h_{2i}, \cdots h_{Mi}\}^T$ is the $M \times 1$ channel vector between the i-th device and M antennas at the BS. Here h_{Mi} denotes the channel coefficient between i-th device and the M-th antenna of the BS, and has a uniform phase. We model the amplitude power of channel $|h_{ji}|^2$ for $j = \{1, 2, \cdots M\}$ as:

$$|h_{ji}|^2 = F_{ji} \cdot GR_i^{-\alpha} \cdot 10^{\frac{S_i}{10}}, \quad i = \{1, 2, \cdots N\} \tag{4.23}$$

where F_{ji} models the short-term Rayleigh fading channel between the i-th device and the j-th antenna, R_i is the distance from the i-th device to the BS, α is the path loss coefficient, and the term G is the normalizing factor for the path loss. The term $S_i \sim \mathcal{N}(0, \sigma^2)$ is normal such that $10^{\frac{S_i}{10}}$ is log-normally distributed and models shadowing behavior. The parameter σ is known as the dB spread or the shadowing factor.

Since the long term path loss dominates the short term fading, and over longer time scales Rayleigh fading is averaged out, we can represent (4.23) as normally distributed by taking the logarithm of (4.23):

$$10\log_{10}|h_{ji}|^2 \text{ s.t. } X_i \sim \mathcal{N}(10\log_{10} R_i^{-\alpha} F_i + 10\log_{10} G, \sigma^2) \tag{4.24}$$

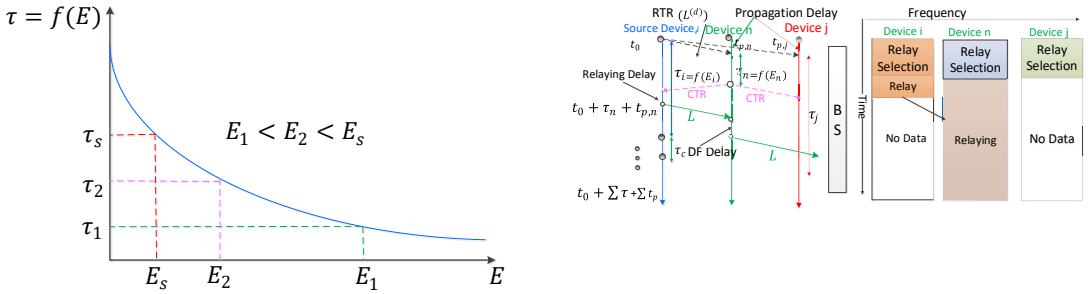
Indeed, a generalized distribution of $|h_{ji}|^2$ can be obtained by considering the combined distribution of S_i,F_{ji}, and R_i, which may become intractable for performance analysis.

If the direct transmission is not energy-efficient (e.g. due to shadowing effect between devices and the BS), the single-antenna source device sends data to a single-antenna relay device using the D2D communication. The received signal at the n-th relay device is given as

$$y_n^{(\mathrm{d})} = \sqrt{P}h_i^{(\mathrm{d})}x_i + v \tag{4.25}$$

where $h_i^{(\mathrm{d})}$ is the fading channel between the i-th source device and the selected relay device n, and v is the AWGN with power N_0. Since the quality of signal received at the neighboring relay can be high, a DF protocol can be used at the relay to transmit the data from the source device to the BS. It is noted that all devices use different RBs separated in time and frequency, and thus, there is no interference even if a single relay device receives signal from multiple source devices as these are sent at different RBs.

For D2D links, we ignore the shadowing effect, similar to [58], [59] [60]. This assumption is justified since the two devices communicate with each other under close proximity



(a) (b)

Figure 4.13. ODSR for three devices with transmission energy $E_1 < E_2 < E_s$ (a) function $f(E)$ (b) timing diagram and resource block allocation.

as per the Long Term Evolution (LTE) standard [140]. We assume that the short-term fading amplitude $|h_i^{(\mathrm{d})}|$ between the i-th source device and the relay device is Rayleigh distributed such that

$$|h_i^{(\mathrm{d})}|^2 = r_i^{-\alpha^{(\mathrm{d})}} F_i^{(d)} \tag{4.26}$$

where $F_i^{(d)}$ follows the exponential distribution, r_i is the distance from the i-th source device to the selected relay device, and $\alpha^{(\mathrm{d})}$ is the path loss exponent between them. Since devices are close each other in D2D communication, the probability that relay devices receive signal at a very high SNR is high, and thus, consume negligible energy compared with the direct transmission.

4.3.2 ODSR Relaying Scheme

In this section, we describe the ODSR, which minimizes energy consumption for data transmission and its distributed implementation based on the timer-based protocol [47].

Criteria of Relaying Device Selection

We consider transmissions of packets with a fixed length of L bits by the source device to the BS in each transmission slot. We assume that all devices transmit with equal power P, and denote the circuit power by P_i^{ckt} for the i-th device. Since the power dissipated in the transmitter and receiver circuits is different for different devices, we consider that the circuit power transmission of the devices is uniformly distributed between $P_{\min}^{\mathrm{ckt}}$ and $P_{\max}^{\mathrm{ckt}}$. Using (4.22), the energy consumed by the i-th source device to transmit its data directly to the BS is:

$$E_i = (P + P_i^{\mathrm{ckt}}) \cdot \frac{L}{B \log_2(1+\gamma_i)} = \frac{\eta_1}{10 \log_{10}(1+\gamma_i)} + \frac{\eta_2 P_i^{\mathrm{ckt}}}{10 \log_{10}(1+\gamma_i)} \tag{4.27}$$

where B is the transmission channel bandwidth, $\eta_1 = 10\log_{10}(2)PL/B$, $\eta_2 = \eta_1/P$, and $\gamma_i = \frac{\sum_{j=1}^{M}|h_{ji}|^2P}{N_0}$ is the received SNR at the BS due to the linear combination of M signals when the signal is transmitted from the i-th device.

Using (4.25), the energy consumed by the D2D communication to relay a data of L bits is:

$$E_i^{(\mathrm{d})} = \frac{\eta_1^{(\mathrm{d})}}{\log(1+\gamma_i^{(\mathrm{d})})} + \frac{\eta_2^{(\mathrm{d})}P_i^{\mathrm{ckt}}}{\log(1+\gamma_i^{(\mathrm{d})})} \tag{4.28}$$

where $\eta_1^{(\mathrm{d})} = \log(2)P^{(\mathrm{d})}L/B$, $\eta_2^{(\mathrm{d})} = \eta_1^{(\mathrm{d})}/P^{(\mathrm{d})}$, and $\gamma_i^{(\mathrm{d})} = \frac{|h_i^{(\mathrm{d})}|^2P^{(\mathrm{d})}}{N_0^{(\mathrm{d})}}$ is the SNR at the relay device when the signal is transmitted at a power $P^{(d)}$ from the i-th source device.

The relay selection criteria for the ODSR is based on the minimum consumed energy for transmission of packet data to the BS as:

$$n = \underset{1\le i\le N}{\operatorname{argmin}}\{E_i\}. \tag{4.29}$$

It is noted that ODSR relay selection requires only the channel information from devices to the BS. It should be noted that the component of the relaying energy $E_i^{(\mathrm{d})}$ is ignored in the relay selection since this may require the Channel State Information (CSI) between the source to relaying devices. In general, the energy consumption of the D2D relaying (due to the close proximity) is lower than the energy consumed in forwarding the data to the BS (which can be affected by the shadow fading) in the second hop, and thus may not affect the relay selection process. It is good to note that we have included $E_i^{(\mathrm{d})}$ while deriving bounds on the energy consumption performance of the ODSR.

There is no advantage of considering circuit power transmission for relay selection if it is assumed equal for all devices (i.e, $P_i^{\mathrm{ckt}} = P^{\mathrm{ckt}}, \forall i$). However, in practice, the circuit transmission power for all devices may not be equal due to different types and specifications of devices in a network. This will lead to a randomness in the circuit power transmissions and the second term in (4.27) will become the ratio of random variables. Under this condition, the relay selection will depend on the circuit transmission power of devices, and analyzing the average energy consumption will be challenging due to an additional term of the ratio of random variables.

4.3.3 Distributed Implementation of ODSR

Distributed implementation of the protocol is desired since the centralized relay selection requires the global information of the CSI. Further, the centralized implementation consumes a large energy overhead due to control signaling. In [47]the authors describe a timer-based distributed protocol for relay selection (controlled by the BS with Request To Send (RTS) and Clear To Send (CTS) signals using instantaneous channel information of both hops. This technique has been found to be useful in many relaying based networks [141, 142]. The authors in [141] have used the protocol of [47] for relay selection using power control at each relays for an energy-efficient transmission.

The distributed implementation of the ODSR is based on the back-off principle of the Carrier Sensing multiple Access (CSMA) in the MAC layer supported with the transmission energy from the physical layer. We define an increasing function $f(E)$ designed judiciously (see Figure 4.13a) such that back-off time $\tau_i = f(E_i), i = 1, \cdots, N$ of the devices has distinct energy index $E_i, i = 1, \cdots, N$. Thus, the considered implementation is based on the criteria of consumed energy with proper adaptations for uplink data transmissions in a wireless network using D2D relaying, as described in the following steps (see Figure 4.13b):

Request To Relaying (RTR)

First, the i-th source device sets its back-off time to $\tau_i = f(E_i)$ and broadcasts an Request To Relaying (RTR) message (with fields such as user ID) to be received by the devices in close proximity. All the devices are capable of decoding the RTR message with the CSI estimated using the RTR message. The CSI is available if devices are already in the discovery mode compliant with the proximity services of LTE [140]. The RTR transmission costs an energy consumption $E_{\text{tx}}^{\text{RTR}}$ to the source device. The energy overhead in decoding the RTR per device is $E_{\text{rx}}^{\text{RTR}}$.

The source device waits for a reply from a potential relay for a duration of $\tau_i + \tau_{\text{c}}$, where τ_{c} is an additional delay to compensate for the propagation delays in D2D communication. This delay corresponds to relay selection overhead, as depicted in Figure 4.13b. If the device does not receive a reply from any device for relaying in the time limit of $\tau_i + \tau_{\text{c}}$, it directly transmits to the BS (step 4), otherwise the data is transmitted through a relay. Note that an increase in the transmission delay is compensated by the use of relay with the best channel which reduces time to transmit the data to the BS.

Distributed Relay Selection

Upon the receipt of a RTR message from the source, each device sets its back-off time to $\tau_j = f(E_j), j \cdots N - 1$. In the opportunistic relaying scheme, the n-th device selected using the criteria in (4.29) has the lowest back-off time, and hence occupies the channel first by responding to the source with a Clear To Relay (CTR) message after a waiting period $\tau_n < \tau_j, n \neq j$. It should be noted that the probability that two users have equal back-off time is zero [47]. Once the selected device transmits the CTR message to the source, all other devices overhear the CTR message (or just a busy tone), and quit the process of relay selection for the given request from the i-th source device. The overhead energies for a response from the relay device are: transmission of CTR message $E_{\text{tx}}^{\text{CTR}}$ and reception of CTR message $E_{\text{tx}}^{\text{CTR}}$.

Source to Relay Transmission

Upon the successful decoding of the CTR message, the source device sends the data packet to the selected relay device with a transmit energy cost E_i^{d} as computed in (4.28). Using the DF protocol, the selected relay device decodes the data from the source device,

encodes it, and transmits to the BS. The DF protocol requires the CSI at the relay device. This can be estimated using the RTR message from the source device after the decision on relay selection. The energy overhead at this stage is: CSI estimation energy E^{CSI}, transmit energy cost E_i^{d}, decoding energy E^{DEC}, and encoding energy E^{ENC}.

Data Transmission

Finally, transmission of data is accomplished by direct transmission from the source or the relay device. The energy consumption in this phase is E_i as computed in (4.27). Note that if a single device happens to act as the source for its data and as the relay for other sources, the data transmission can be done simultaneously using full-duplexing mode.

In the following sections, we analyze the performance of the opportunistic relaying by deriving bounds on the average energy consumption by the devices for data transmission.

4.3.4 Performance Bounds of ODSR

Given the steps of distributed relaying described in the subsection 4.3.3, the total consumed energy by the ODSR is:

$$E^{\mathrm{TOTAL}} = p(E_{\mathrm{ov}}^{\mathrm{RELAY}} + E^{\mathrm{D2D}} + E^{\mathrm{RELAY}}) + (1-p)(E^{\mathrm{DT}} + E_{\mathrm{ov}}^{\mathrm{DT}}) \tag{4.30}$$

where p is the probability of the relay-assisted data transmission i.e., $p = Pr(E^{\mathrm{RELAY}} + E^{\mathrm{D2D}} < E^{\mathrm{DT}})$. We denote E^{RELAY} as the energy consumed by the selected relay to transmit the data packet to the BS and E^{D2D} as the transmission energy by the source device to the selected relay. Further, $E_{\mathrm{ov}}^{\mathrm{RELAY}} = E_{\mathrm{tx}}^{\mathrm{RTR}} + (N-1)E_{\mathrm{rx}}^{\mathrm{RTR}} + E_{\mathrm{tx}}^{\mathrm{CTR}} + E_{\mathrm{rx}}^{\mathrm{CTR}} + E^{\mathrm{CSI}} + E^{\mathrm{DEC}} + E^{\mathrm{ENC}}$ is the overhead energy required for relay selection in the case of D2D communication, E^{DT} denotes the energy consumed for data transmission directly to the BS when the direct transmission is found to be more energy-efficient than the relay-assisted transmission, and $E_{\mathrm{ov}}^{\mathrm{DT}} = E_{\mathrm{tx}}^{\mathrm{RTR}} + (N-1)E_{\mathrm{rx}}^{\mathrm{RTR}}$ is overhead energy for the relay selection.

It is noted that the direct transmission (i.e., without relaying protocol) does not incur any overhead energies. However, the overhead energy $E_{\mathrm{ov}}^{\mathrm{RELAY}}$ of the ODSR is also low (see Table 4.1 signaling involved is very short and the signaling messages are sent to other local devices with very low power. This is illustrated through simulations in realistic scenarios of a wireless network in Section 4.3.5.

Average Energy Consumption of D2D Transmission: $\bar{E}^{\mathrm{D2D}}$

In this subsection, we analyze the overhead energy of the ODSR due to the D2D transmission. Under the Rayleigh fading for the D2D channel, the SNR $\gamma^{(d)}$ as given in (4.28) (we drop the index i)) is exponential distributed with probability distribution function Probability Distribution Function (PDF) $f(\gamma^{(\mathrm{d})}) = \frac{1}{\bar{\gamma}^{(\mathrm{d})}} \mathrm{e}^{-\gamma^{(\mathrm{d})}/\bar{\gamma}^{(\mathrm{d})}}$ where $\bar{\gamma}^{(\mathrm{d})} =$

$\mathbb{E}[\gamma^{(d)}] = \int_0^\infty \gamma^{(d)} f(\gamma^{(\mathrm{d})}) d\gamma^{(d)}$ is the average SNR. Using (4.28), the average consumed energy for the D2D relaying:

$$\bar{E}^{\mathrm{D2D}} = \Big(\eta_1^{(\mathrm{d})} + \eta_2^{(\mathrm{d})}\mathbb{E}[P^{\mathrm{ckt}}]\Big) \times \frac{1}{\bar{\gamma}^{(\mathrm{d})}} \int_{\gamma_{\mathrm{th}}^{(\mathrm{d})}}^{\infty} \frac{1}{\log(1+x)} e^{-x/\bar{\gamma}^{(\mathrm{d})}} \mathrm{d}x \tag{4.31}$$

where $\gamma_{\mathrm{th}}^{(\mathrm{d})}$ is the threshold SNR (in linear scale) for the D2D communication. Using the series expansion of exponential function in (4.31), we get an exact expression of the expected energy consumption for the D2D relaying:

$$\bar{E}^{\mathrm{D2D}} = \frac{1}{\bar{\gamma}^{(\mathrm{d})}}(\eta_1^{\mathrm{d}} + 0.5\eta_2^{(\mathrm{d})}(P_{\max}^{\mathrm{ckt}} + P_{\min}^{\mathrm{ckt}})) \times \sum_{k=0}^{\infty} \frac{(-1)^k}{k!} \frac{1}{(\bar{\gamma}^{(\mathrm{d})})^k} [E_i(\gamma_{\max} + k\gamma_{\max}) - E_i(\gamma_{\mathrm{th}}^{(\mathrm{d})} + k\gamma_{\mathrm{th}}^{(\mathrm{d})})]. \tag{4.32}$$

Further, we provide simple bounds on (4.31) in the following Theorem:

Theorem 1. *If $P_{\min}^{\mathrm{ckt}}$ and $P_{\max}^{\mathrm{ckt}}$ are minimum and maximum circuit transmit power of all devices, respectively, γ_{th} is the threshold SNR, and $\eta_1^{(\mathrm{d})} = 10\log(2)P^{(\mathrm{d})}L/B$, $\eta_2^{(\mathrm{d})} = \eta_1^{(\mathrm{d})}/P^{(\mathrm{d})}$, then the expected energy consumption for D2D under Rayleigh fading channel with average SNR $\bar{\gamma}^{(\mathrm{d})}$ is bounded as:*

$$\begin{aligned}(\eta_1^{(\mathrm{d})} + 0.5\eta_2^{(\mathrm{d})}(P_{\max}^{\mathrm{ckt}} + P_{\min}^{\mathrm{ckt}})) \times \Big(\frac{1}{\bar{\gamma}^{(\mathrm{d})}} \log_e(1 + \frac{\bar{\gamma}^{(\mathrm{d})}}{\gamma_{\mathrm{th}}^{(\mathrm{d})}}) - \frac{1}{(\bar{\gamma}^{(\mathrm{d})})^2} \log(1 + \frac{\bar{\gamma}^{(\mathrm{d})}}{\gamma_{\mathrm{th}}^{(\mathrm{d})}})\Big) \leq \bar{E}^{\mathrm{D2D}} \\ \leq (\eta_1^{(\mathrm{d})} + 0.5\eta_2^{(\mathrm{d})}(P_{\max}^{\mathrm{ckt}} + P_{\min}^{\mathrm{ckt}})) \times \Big(\frac{\bar{\gamma}^{(\mathrm{d})}}{\bar{\gamma}^{(\mathrm{d})} + \gamma_{\mathrm{th}}^{(\mathrm{d})}} + \frac{1}{\bar{\gamma}^{(\mathrm{d})} + \gamma_{\mathrm{th}}^{(\mathrm{d})}} \log(1 + \frac{\bar{\gamma}^{(\mathrm{d})}}{\gamma_{\mathrm{th}}^{(\mathrm{d})}})\Big)\end{aligned} \tag{4.33}$$

Proof. Using the expectation of uniform random variable and applying logarithm inequality $\frac{x}{x+1} \leq \log(1+x) \leq x$ [143], the integral in (4.31) for expected energy in D2D relaying can be represented in terms of exponential integral:

$$\begin{aligned}(\eta_1^{(\mathrm{d})} + 0.5\eta_2^{(\mathrm{d})}(P_{\max}^{\mathrm{ckt}} + P_{\min}^{\mathrm{ckt}})) \frac{1}{\bar{\gamma}^{\mathrm{d}}} E_1(\frac{\gamma_{\mathrm{th}}^{(\mathrm{d})}}{\bar{\gamma}^{d}}) \leq \bar{E}^{\mathrm{D2D}} \leq \\ (\eta_1^{(\mathrm{d})} + 0.5\eta_2^{(\mathrm{d})}(P_{\max}^{\mathrm{ckt}} + P_{\min}^{\mathrm{ckt}}))\Big(\exp(-\frac{\gamma_{\mathrm{th}}^{(\mathrm{d})}}{\gamma^{\mathrm{d}}}) + \frac{1}{\bar{\gamma}^{(\mathrm{d})}} E_1(\frac{\gamma_{\mathrm{th}}^{(\mathrm{d})}}{\bar{\gamma}^{(\mathrm{d})}})\Big)\end{aligned} \tag{4.34}$$

Further, we use the inequality on exponential integral $0.5\exp(-x)\log(1+2/x) < E_1(x) < \exp(-x)\log(1+1/x)$ and $\exp(x) > 1+x$ to get (4.33) of Theorem 1.

□

From (4.32) and (4.33), it can be seen that the expected energy decreases with an increase in the average SNR at the relaying device. Since the relay devices have a higher average SNR due to proximity with the source device in the D2D communication, the

energy overhead of the relaying among devices is negligible as compared with the transmission of data to the BS.

Average Energy Consumption without Relaying: $\bar{E}^{\mathrm{DT}}$

We derive an expression on the expected consumed energy without D2D relaying (i.e., direct transmission). Each device transmits its data to the BS, if $E^{\mathrm{RELAY}} + E^{\mathrm{D2D}} \geq E^{\mathrm{DT}}$. Using a simple inequality, $10\log_{10}(z) \leq 10\log_{10}(1+z) \leq 1 + 10\log_{10}(z), z \neq 0$ in (4.27), we get bounds on the energy consumption of a device (we drop the index i) for the direct transmission as

$$\frac{\eta_1 + \eta_2 P^{\mathrm{ckt}}}{1+X} \leq E^{\mathrm{DT}} \leq \frac{\eta_1 + \eta_2 P^{\mathrm{ckt}}}{X} \tag{4.35}$$

where $X = 10\log_{10}(\gamma)$. The term $\sum_{j=1}^{M} |h_{ji}|^2$ in $\gamma_i = \frac{\sum_{j=1}^{M} |h_{ji}|^2 P}{N_0}$ can be approximated as lognormal distributed since $|h_{ji}|^2$ is lognormal (see (4.24)) and sum of log-normal random variables can also be approximated as log-normal [144]. Moreover, each antenna gets the same shadowing effect as is typical in wireless channel models [145]. Thus γ is log-normal distributed with a spreading parameter σ^2 in dB, $X \sim \mathcal{N}(\bar{\gamma}, \sigma^2)$ with

$$\bar{\gamma} \;\; = 10\log_{10} M + 10\log_{10} F + 10\log_{10} R^{-\alpha} + 10\log_{10} G + 10\log_{10} P/N_0$$

Considering different specifications of user devices in a network, the devices can have A different circuit power consumption models. Thus, we model the circuit power to be uniformly distributed between $P^{\mathrm{ckt}}_{\mathrm{min}}$ and $P^{\mathrm{ckt}}_{\mathrm{max}}$ representing minimum and maximum circuit transmit powers, respectively.

Taking expectation in (4.35) and noting the independence between the numerator and denominator terms, we get an upper bound on the expected energy consumption with direct transmission as:

$$\bar{E}^{\mathrm{DT}} \leq \mathbb{E}[\eta_1 + \eta_2 P^{\mathrm{ckt}}]\mathbb{E}[\frac{1}{X}] = \left(\eta_1 + \eta_2 \mathbb{E}[P^{\mathrm{ckt}}]\right) \frac{1}{\sqrt{2\pi}\sigma} \int_{\gamma_{\mathrm{th}}}^{\infty} \frac{1}{x} e^{-\frac{(x-\bar{\gamma})^2}{2\sigma^2}} \mathrm{d}x \tag{4.36}$$

where γ_{th} in dB is a SNR threshold. The threshold SNR is selected to achieve a minimum data rate requirement below which communication is possible. The expectation has been taken over SNR γ. A lower bound can be similarly obtained by replacing $\bar{\gamma}$ with $\bar{\gamma} + 1$.

Theorem 2. *If $P^{\mathrm{ckt}}_{\mathrm{min}}$ and $P^{\mathrm{ckt}}_{\mathrm{max}}$ are minimum and maximum circuit transmit power of all devices, respectively, γ_{th} is the threshold SNR in dB, and $\eta_1 = 10\log_{10}(2)PL/B$, $\eta_2 = \eta_1/P$, then the expected energy with the direct transmission in a log-normal fading channel with average SNR $\bar{\gamma}$ and variation σ (in dB) is bounded as:*

$$\frac{(\eta_1 + 0.5\eta_2(P^{\mathrm{ckt}}_{\mathrm{max}} + P^{\mathrm{ckt}}_{\mathrm{min}}))}{(\bar{\gamma}+1)} \exp\left(\frac{\sigma^2}{2(\bar{\gamma}+1)^2}\right) \times Q\left(\frac{\sigma}{(\bar{\gamma}+1)} + \frac{(\gamma_{\mathrm{th}} - \bar{\gamma} - 1)}{\sigma}\right) \leq \bar{E}^{\mathrm{DT}}$$
$$\leq (\eta_1 + 0.5\eta_2(P^{\mathrm{ckt}}_{\mathrm{max}} + P^{\mathrm{ckt}}_{\mathrm{min}}))[\mathcal{I}_1^{\mathrm{DT}}(\bar{\gamma}, \sigma) + \mathcal{I}_2^{\mathrm{DT}}(\bar{\gamma}, \sigma)]$$

where

$$\mathcal{I}_1^{\mathrm{DT}}(\bar{\gamma},\sigma)=\frac{\sigma}{\sqrt{2\pi}(2\sigma^2+\bar{\gamma}^2)}\Big[2\sqrt{2}\sigma\log\left(\frac{\bar{\gamma}}{\gamma_{\mathrm{th}}}\right)\times\log\left(1+\left(\frac{\bar{\gamma}-\gamma_{\mathrm{th}}}{\sqrt{2}\sigma}\right)^2\right)+\arctan\left(\frac{\bar{\gamma}-\gamma_{\mathrm{th}}}{\sqrt{2}\sigma}\right)^2\Big] \tag{4.37}$$

$$\mathcal{I}_2^{\bar{\gamma},\mathrm{DT}}(\sigma)=\frac{\exp[-\bar{\gamma}^2/2\sigma^2]}{4\sqrt{2\pi}\sigma}\Big[2\pi\mathrm{erfi}\left(\frac{\bar{\gamma}}{\sqrt{2}\sigma}\right)-2E_1\left(\frac{\bar{\gamma}^2}{2\sigma^2}\right)+\log\left(\frac{\bar{\gamma}^2}{2\sigma^2}\right)+4\log\left(\frac{\sqrt{2}\sigma}{\bar{\gamma}}\right)-\log\left(\frac{\sigma^2}{\bar{\gamma}}\right)\Big] \tag{4.38}$$

Proof. The integral in (4.36) can be represented as a sum of two integrals:

$$\mathcal{I}_{\mathrm{ub}}=\frac{1}{\sqrt{\pi}}\left[\int_0^{\frac{\bar{\gamma}-\gamma_{\mathrm{th}}}{\sigma\sqrt{2}}}\frac{1}{\bar{\gamma}-\sqrt{2}t\sigma}e^{-t^2}\mathrm{d}t+\int_0^{\infty}\frac{1}{\bar{\gamma}+\sqrt{2}t\sigma}e^{-t^2}\mathrm{d}t\right]$$

We use the standard mathematical procedure on the second integral in (4.39) to get an exact solution $\mathcal{I}_2^{\mathrm{DT}}(\bar{\gamma},\sigma)$ as given in (4.38). Using $\exp[-x^2]\leq\frac{1}{1+x^2}$ and applying the partial fraction method, an upper bound of the first integral is given as $\mathcal{I}_1^{\mathrm{DT}}(\bar{\gamma},\sigma)$. This has been presented in (4.38). Using these, and the average of uniform random variable, we get the upper bound (4.37) of Theorem 2. For the lower bound, we use (4.35) and $1+z\leq e^z$ to get the first integral of (4.39) as

$$\mathcal{I}_{\mathrm{lb}}=\frac{1}{\sqrt{2\pi}(\bar{\gamma}+1)}\int_{\frac{(\gamma_{\mathrm{th}}-\bar{\gamma}-1)}{\sigma}}^{\infty}e^{-\frac{x^2}{2}-\frac{\sigma}{\bar{\gamma}+1}x}\mathrm{d}x \tag{4.39}$$

Completing the expression in the exponential function in a square form and representing the integral into Gaussian Q-function with a simple substitution, we get the lower bound (4.37) of Theorem 2.

□

The derived bounds in (4.37) are presented in terms of simple mathematical functions. It can be seen that a lower average SNR increases the energy consumption for the direct transmission, thus necessitating the use of relaying.

Average Energy Consumption with Relaying: $\bar{E}^{\mathrm{RELAY}}$

Now, we derive an expression for the average energy consumed $\bar{E}^{\mathrm{RELAY}}$ by the device to the BS in log-normal fading with the selection criteria defined in (4.29). To simplify the model, we assume that the relaying devices are in the vicinity of the source, so that the path loss of all possible relays are similar [146], but spread enough to experience independent shadowing. We also assume the circuit power is the same for each device i.e., $P_{\mathrm{min}}^{\mathrm{ckt}}=P_{\mathrm{max}}^{\mathrm{ckt}}=P^{\mathrm{ckt}}$. Using the selection criteria in (4.29) for the log-normal shadowing

in (4.35), we get:

$$E^{\text{RELAY}} \leq \frac{\eta_1 + \eta_2 P^{\text{ckt}}}{X_{(n)}} \tag{4.40}$$

where $X_{(n)} = \max(X_1, X_2, X_3, \cdots, X_N)$ with $X_i = 10\log_{10}(\gamma_i)$, $1 \leq i \leq N$. It follows from order statistics that the CDF of $X_{(n)}$ is given as $F_{X_{(n)}}(x) = [F_X(x)]^N$, where $F_X(x) = [1/2 + 1/2\text{erf}(\frac{x-\bar{\gamma}}{\sqrt{2\sigma^2}})]$ is the CDF of normal distribution. The PDF of $X_{(n)}$ is $f_{X_{(n)}}(x) = N[F_X(x)]^{N-1}[f_X(x)]$ where $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\bar{\gamma})^2}{2\sigma^2}}$ is the PDF of normal distribution.

Thus, the average consumed energy $\bar{E}^{\text{RELAY}} = \mathbb{E}[E^{\text{RELAY}}]$ can be expressed as:

$$\bar{E}^{\text{RELAY}} \leq \mathbb{E}[\eta_1 + \eta_2 P^{\text{ckt}}] \int_{\gamma_{\text{th}}}^{\infty} \frac{N}{x} [F_X(x)]^{N-1} [f_X(x)] \mathrm{d}x \tag{4.41}$$

Using the integration by parts and $F_X(x) = Q(\frac{\bar{\gamma}-\gamma_{\text{th}}}{\sigma})$, we can represent (4.41) as:

$$\bar{E}^{\text{RELAY}} \leq \mathbb{E}[\eta_1 + \eta_2 P^{\text{ckt}}] \times \left(\mathcal{I}_1^{\text{RELAY}}(N,\sigma) + \mathcal{I}_2^{\text{RELAY}}(N,\sigma) - \frac{1}{\gamma_{\text{th}}} Q^N(\frac{\bar{\gamma}-\gamma_{\text{th}}}{\sigma}) \right)$$

where

$$\begin{aligned} \mathcal{I}_1^{\text{RELAY}}(N,\sigma) &= \int_{\frac{\gamma_{\text{th}}-\mu}{\sigma}}^{0} \frac{1}{(x\sigma+\bar{\gamma})^2} (1-Q(x))^N \mathrm{d}x \\ \mathcal{I}_2^{\text{RELAY}}(N,\sigma) &= \int_{0}^{\infty} \frac{1}{(x\sigma+\bar{\gamma})^2} (1-Q(x))^N \mathrm{d}x \end{aligned} \tag{4.42}$$

Theorem 3. *If P^{ckt} is the circuit transmit power of each device, γ_{th} is the threshold SNR in dB, and $\eta_1 = 10\log_{10}(2)PL/B$, $\eta_2 = \eta_1/P$, then the average energy consumption with relaying from N devices in a log-normal fading channel with average SNR $\bar{\gamma}$ and variation σ (in dB) is bounded as:*

$$\bar{E}^{\text{RELAY}} \leq (\eta_1 + \eta_2 P^{\text{ckt}}) \times \left(\mathcal{I}_1^{\text{RELAY}}(N,\sigma) + \mathcal{I}_2{}^{\text{RELAY}}(\text{N},\sigma) - \frac{1}{\gamma_{\text{th}}} Q^N(\frac{\bar{\gamma}-\gamma_{\text{th}}}{\sigma}) \right) \tag{4.43}$$

where $\mathcal{I}_1^{\text{RELAY}}(N,\sigma)$ and $\mathcal{I}_2^{\text{RELAY}}(N,\sigma)$ are given in (4.44) and (4.44) (see next page), respectively.

$$\begin{aligned} \mathcal{I}_1^{\text{RELAY}}(N,\sigma) \leq \frac{\sigma}{(2)^N(2\sigma^2+N\bar{\gamma}^2)^2} \Bigg[& 2\sigma^2 \left(2\sigma^2 + N\bar{\gamma}^2\right) \left(\frac{1}{\gamma_{\text{th}}} - \frac{1}{\bar{\gamma}}\right) + \\ & + 4N\sigma^2\bar{\gamma} \log\left(\frac{\gamma_{\text{th}}}{\bar{\gamma}}\right) + +2N\sigma\mu \log\left(1 + \frac{N}{2}\left(\frac{\bar{\gamma}-\gamma_{\text{th}}}{\sigma}\right)\right) + \\ & + \sqrt{2N}\left(N\bar{\gamma}^2 - 2\sigma^2\right) \arctan\left(\sqrt{\frac{N}{2}}\left(\frac{\bar{\gamma}-\gamma_{\text{th}}}{\sigma}\right)\right) \Bigg] \end{aligned}$$

$$\mathcal{I}_2^{\text{RELAY}}(N,\sigma) \leq \sigma \sum_{r=0}^{N} \binom{\frac{N}{2}}{2r} \frac{1}{4^r} \Psi(r,\sigma,\bar{\gamma}) - \sum_{r=0}^{N} \binom{\frac{N}{2}}{2r+1} [f(\kappa)]^{2r+1} \Psi((2r+1)\kappa,\sigma,\bar{\gamma}), \tag{4.44}$$

where $f(\kappa) = \frac{\exp((\pi(\kappa-1)+2)^{-1})}{2\kappa}\sqrt{\frac{1}{\pi}(\kappa-1)(\pi(\kappa-1)+2)}$, $\kappa \geq 1$, *and function* $\Psi(r,\sigma,\bar{\gamma})$:

$$\begin{aligned}\Psi(N,a,b) = \int_0^{\infty} \frac{\exp[-Nx^2]}{(ax+b)^2}\mathrm{d}x = \frac{1}{2a^3b} e^{-\frac{nb^2}{a^2}} \Bigg(& 2\pi b^2 N \, erfi\left(\frac{b\sqrt{N}}{a}\right) - 2b^2 N Ei\left(\frac{b^2N}{a^2}\right) + \\ & + 2a^2 e^{\frac{b^2N}{a^2}} - 2\sqrt{\pi}ab\sqrt{N}e^{\frac{b^2N}{a^2}} - b^2N\log\left(\frac{a^2}{b^2N}\right) + b^2N\log\left(\frac{b^2N}{a^2}\right) + \\ & + 4b^2N\log\left(\frac{a}{b}\right) - 2b^2N\log(N)\Bigg), N>0, a>0, b>0\end{aligned} \tag{4.45}$$

Proof. An upper bound on $I_1^{\text{RELAY}}(N,\sigma)$ in (4.42) can be obtained using $Q(t) = 1-Q(-t)$ with Chernoff bound $Q(t) \leq \frac{1}{2}\exp[-t^2/2]$, and $\exp[-z] < \frac{1}{1+z}$ to express $I_1^{\text{RELAY}}(N,\sigma)$ as a polynomial function:

$$I_1^{\text{RELAY}}(N,\sigma) \leq \frac{1}{(2)^N} \int_0^{\frac{\bar{\gamma}-\gamma_{\text{th}}}{\sigma}} \frac{1}{(\bar{\gamma}-t\sigma)^2(1+\frac{N}{2}t^2)}\mathrm{d}t \tag{4.46}$$

We use the partial fraction to solve the integral in (4.46) which is given in (4.44). To analyze $I_2^{\text{RELAY}}(N,\sigma)$, we use the binomial expansion of $(1-Q(x))^N$ and interchange the summation and the integration to get

$$\begin{aligned} I_2^{\text{RELAY}}(N,\sigma) &= \sum_{k=0}^{N} \binom{N}{k} (-1)^k \int_0^{\infty} \frac{[Q(x)]^k}{(x\sigma+\bar{\gamma})^2}\mathrm{d}x \\ &= \sum_{r=0}^{N} \binom{N/2}{2r} \int_0^{\infty} \frac{[Q(x)]^{2r}}{(x\sigma+\bar{\gamma})^2}\mathrm{d}x - \sum_{r=0}^{N} \binom{N/2}{2r+1} \int_0^{\infty} \frac{[Q(x)]^{2r+1}}{(x\sigma+\bar{\gamma})^2}\mathrm{d}x \end{aligned}$$

Then, we use Chernoff bounds $f(\kappa)\exp[-\kappa x^2/2] \leq Q(x) \leq \frac{1}{2}\exp[-x^2/2]$, where $f(\kappa) = \frac{\exp((\pi(\kappa-1)+2)^{-1})}{2\kappa}\sqrt{\frac{1}{\pi}(\kappa-1)(\pi(\kappa-1)+2)}, \kappa \geq 1$ [147] appropriately in (4.47) to represent the integral terms in the form $\int_0^{\infty}\frac{\exp[-Nx^2]}{(ax+b)^2}\mathrm{d}x = \Psi(N,a,b)$. Using standard mathematical procedures, closed-form expression of $\Psi(N,a,b)$ is given in (4.45), and thus we get (4.44). This concludes the proof of Theorem. □

While deriving (4.44), we have used Chernoff type of bounds of the Q-function in (4.47). We further simplify the expression $I_2^{\text{RELAY}}(N,\sigma)$ in (4.47) by applying an approximation $Q(x) \approx \exp(q_1x^2+q_2x+q_3)$, where $q_1 = -0.4920, q_2 = -0.2287, q_3 = -1.1893$ [148] to get an approximate expression on $\mathcal{I}_2^{\text{RELAY}}(N,\sigma)$, as presented in Appendix B.

Thus, using results of Theorem 1, Theorem 2, and Theorem 3 in (4.30), we can express the energy consumption performance of the ODSR in terms of known mathematical functions. In what follows, we provide a scaling law on the average energy consumption of the relaying to the number of devices in a network for better insight on the network performance.

Theorem 4. *If P^{ckt} is the circuit transmit power of devices and $\eta_1 = 10\log_{10}(2)PL/B$, $\eta_2 = \eta_1/P$, then the average consumed energy with a single relay selection from N devices in a log-normal shadow fading channel with average SNR $\bar{\gamma}$ and variation σ (in dB) is upper bounded as:*

$$\begin{aligned}\bar{E}^{\text{RELAY}} \leq \left(\eta_1 + \eta_2 P^{\text{ckt}}\right) \times \Big(\frac{1}{2^N}\frac{1}{\gamma_{\text{th}}} + \frac{1}{\sigma}\Big(\frac{1}{\bar{\gamma} + \sigma\sqrt{c_I \log(N)}} \\ + \sum_{i=1}^{I-1}\left(\frac{1}{1+\kappa_2 N^{(1-c_i)}}\right)\left(\frac{1}{\bar{\gamma} + \sigma\sqrt{c_{i-1}\log(N)}}\right)\Big)\end{aligned} \tag{4.47}$$

where I is a positive integer, $\kappa_2 = 0.3885$ is a constant, and $0 \leq c_i \leq 1$, $c_0 = 0$, $i = 1, 2, \cdots I$. Further, energy consumption scales as

$$\bar{E}^{\text{RELAY}} = \mathcal{O}\left(\frac{\eta_1 + \eta_2 P^{\text{ckt}}}{\bar{\gamma} + \sigma\sqrt{c_I \log(N)}}\right) \tag{4.48}$$

where $0 \leq c_I \leq 1$.

Proof. The proof is presented in Appendix C. □

From the scaling law in (4.48), it can be seen that energy consumption reduces logarithmic with the number of devices. Hence, near-optimal performance can be achieved with only a few nearby devices selected for D2D relaying. This reduces latency and energy overhead in large scale networks.

4.3.5 Simulation and Numerical Analysis

This section demonstrates the energy consumption performance of the ODSR through numerical analysis and simulations using MATLAB software. We compare the ODSR performance with the optimal and no-relaying (denoted by "direct") schemes. The optimal criteria is based on the relay selection considering energy consumed in both the hops. We use the energy model presented in [149] to compute the energy consumption by the devices for data transmission. We have considered channel models from 3GPP and 5G channels for our simulations [150, 151].

Direct Transmission versus Relaying

First, we demonstrate the energy consumption performance of relaying by considering various path loss configurations and multi-path fading from 3GPP 5G wireless channel

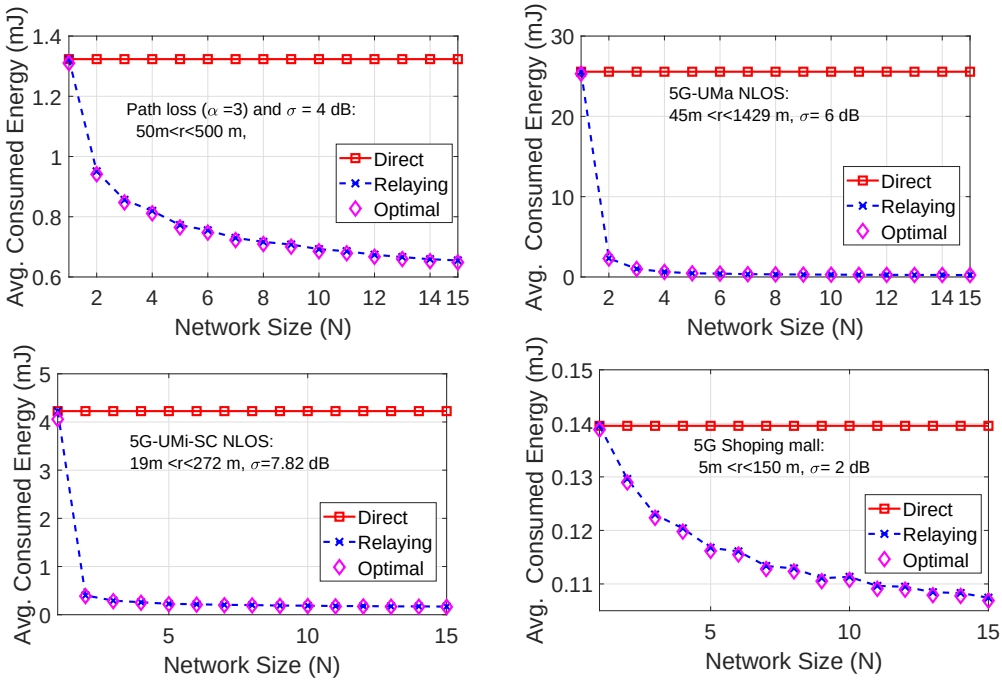


Figure 4.14. Energy consumption performance of opportunistic relaying compared to direct transmission over wireless fading channels for various network scenarios. Different acronyms are UMa: Urban macro, UMi: Urban micro, SC: Street Canyon, NLOS: non-line of sight.

models, as shown Figure 4.14. The log-normal spreading factor ranges from 2 dB to 7.8 dB. We consider short term fading using the Tapped Delay Line Type A (TDL-A) model with delay spread 100 ns [151]. The channel bandwidth is 720 KHz, and the carrier center frequency is 6 GHz. The background noise for each device and the BS is taken as -174 dBm/HZ with a noise figure of 5 dB. It can be seen from Figure 4.14 that the relaying achieves significant improvement compared to the direct transmission for various wireless channels when the shadowing effect is dominant. However, when the the shadowing is minimal (i.e. $\sigma = 2$ dB), the relaying performs very similar to the direct transmission. This motivates us to use relaying based techniques for data transmissions over strong shadow fading channels. The simulation results also show a near-optimal performance of the proposed relaying scheme.

ODSR Performance

In order to demonstrate the ODSR performance, we emulate a wireless network using the 3GPP WINNER II wireless fading model and simulation parameters in line with 3GPP recommendations [150]. This simulation environment enables us to include the overhead energy consumed by the control signaling for a fair comparison with the no-relaying and optimal schemes. For each transmission, a data packet length of $L = 1024$ bytes is considered, and the size of D2D request/reply data is $L^{(d)} = 10$ bytes.

Table 4.1. Average energy consumption (in μJ) of various overheads obtained using simulation under 3GPP model.

$\bar{E}_{\text{tx}}^{\text{RTR}}$	$\bar{E}_{\text{rx}}^{\text{RTR}}$	$\bar{E}_{\text{tx}}^{\text{CTR}}$	$\bar{E}_{\text{rx}}^{\text{CTR}}$	$\bar{E}_{\text{tx}}^{\text{D2D}}$	$\bar{E}_{\text{rx}}^{\text{D2D}}$
11.60	4.50	3.35	1.30	350.5	135.4

The channel model considers all three losses: path-loss, short-term fading, and long-term shadowing. The fading channel between the device and the BS is urban macro log-normal shadowing (spreading factor $\sigma = 4$ dB) while the channel between devices is modeled as Rayleigh fading generated by the extended pedestrian A (EPA) with 9 random taps [152]. The devices are assumed to be moving at a speed of 3 km/h. We consider a single-cell network with up to 150 devices distributed uniformly in a radius of 50m to 500 m with a BS in the center. The background noise for each device and the BS is taken as -174 dBm/HZ. We consider 20 dB of interference at the BS due to inter-cell interference coming from base stations of adjacent cells. We assume transmission power 23 dBm, transmission bandwidth 200 KHz, and initial energy 0.72mWh for all devices. We assume that the communication range for the D2D relaying is within 50 m.

In Table 4.1, we present the components of average consumed energy for various overheads. This can be considered negligible by comparing the energy required for data transmission.

In Figure 4.15, we analyze the performance of ODSR in terms of average energy consumption, network energy efficiency, and the lifetime of the network. The energy efficiency (bits per Joule) of the network is computed as the ratio of channel capacity of all the nodes to the total power consumption (including the circuit power) of the network. We define the lifetime of the network by the average number of transmissions before the battery of the first device of the network is depleted. The figures show that the relaying provides significant performance improvement comparing to the no-relaying scheme. Further, the ODSR achieves the near-optimal performance with only a few relaying devices i.e., within $N = 25$. This happens because the log-normal shadowing of the second hop provides sufficient diversity to achieve the near-optimal performance with a few relaying devices. However, there is a loss in the average number of transmissions by the ODSR compared to the optimal, as shown in Figure 4.15c. This is due to the fact that an incremental decrease in the consumed energy results in a higher cumulative gain in the average number of transmissions.

Scaling Law

Finally, we verify the analytical bounds and the scaling law derived in this section by considering a transmission model without overhead energies, as depicted in Figure 4.15a. We consider a network of 10 to 10^5 devices situated uniformly at 300 m from the BS, situated in the center. For each transmission, a packet length of $L = 2$ MB is considered for a faster simulation in a large network. We consider channel between devices to the

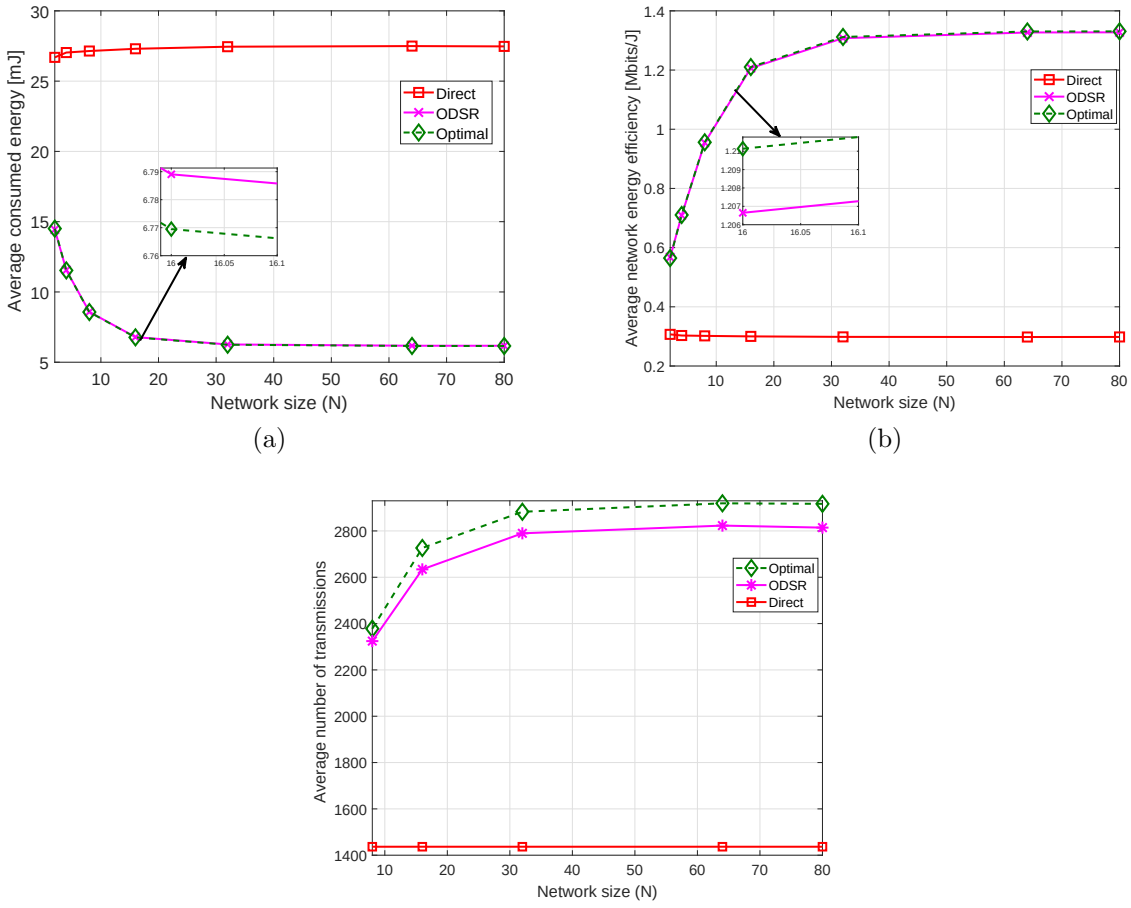
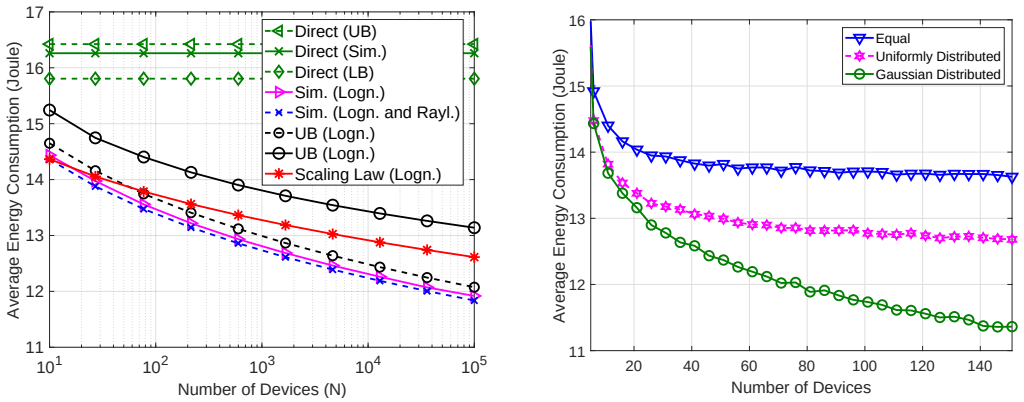


Figure 4.15. Performance of ODSR comparing with the optimal and no-relaying schemes under 3GPP WINNER II fading channels.

BS to be log-normal distributed with a spreading factor of 4 dB and a path loss exponent $\alpha = 4$. The channel between devices is assumed to be Rayleigh fading with a path loss exponent $\alpha = 3$. The transmit power for each device is set to 23 dBm. For scaling law verification, we consider $M = 4$, $c_M = 0.99$, $\delta_M = \ln(N)$, $\delta_1 = \delta_M/4$, $\delta_2 = \delta_M/2$ and $\delta_3 = 3\delta_M/4$ based on Theorem 4.

It can be seen from Figure 4.16a that the short-term fading has a negligible impact on the energy consumption compared to the long-term shadowing effect. Moreover, the figure verifies the analytical bounds and the scaling law on the average consumed energy. It can also be seen that the energy consumption reduces logarithmically with the number of devices. We have also validated bounds of average energy consumption for the direct transmission (Theorem 2) and relayed transmission (as given in Theorem 3 and Theorem 4) with the simulation results.

To verify the effect of randomness of the circuit power transmissions on the relay selection, we assume two probability distribution functions: uniformly distributed between $0.5P^{\text{ckt}}$ and $1.5P^{\text{ckt}}$ and Gaussian distribution $N \sim (P^{\text{ckt}}, 0.03P^{\text{ckt}})$. Figure 4.16b shows



(a) Verification of scaling law and bounds.

(b) Effect of distribution of circuit transmission power.

Figure 4.16. Validation of derived analytical bounds and effect of circuit transmission power on the relaying performance.

that the relay selection depends on the distribution of circuit transmission power of devices. Moreover, the average energy consumption impact is more pronounced when the randomness in the circuit transmission power is high.

4.3.6 Conclusion

In this section we have analyzed the energy consumption performance of a D2D based opportunistic relaying scheme for uplink data transmissions in a mobile network. Furthermore, we have derived closed-form expressions and analytical bounds of the considered ODSR scheme under log-normal shadowing. The analytical expressions show that the ODSR achieves significant performance gain when the devices are in heavy shadowing area with respect to the BS while the devices enjoy high quality channel for inter-user D2D communication with negligible energy overhead. Moreover, the derived scaling law on the consumed energy shows that a near-optimal performance can be achieved in log-normal shadowing with a few devices. This reduces the latency and overhead energy consumed by the devices in the selection of the relays. By considering several realistic mobile network environments, we have shown that the ODSR achieves a near-optimal performance using only few devices in the network. This can be useful to reduce latency and overhead energy consumption in a large scale network. As such, the ODSR achieves an approximately 300% decrease in energy consumption using only 16 relaying devices compared to the direct transmissions. This significant reduction in energy consumption will increase the life time of the network for ubiquitous communications under fading channels. This section completes a solution for collecting of the mobile network information from the mobile devices, represented by the UEs, sensors, vehicles, etc.

4.4 Increasing number of communicating users beyond 5G

The collected information about the UEs provides necessary data to optimize the mobile networks. One of the ways to improve the performance of the mobile networks in terms of the UEs satisfaction with the provided data rates is to deploy the FlyBSs. For the deployment of the FlyBSs it is necessary to have the mobile network information from the UEs, as described in the previous sections. The deployment of the FlyBSs leads to an improved UE channel quality, i.e., higher SINR, due to lower distance between the UE and the FlyBS and higher LOS probability [28]. However, in case, when the FlyBSs operate on the same frequency as the deployed BSs (scenario where the spectral efficiency is the highest), the benefit of the FlyBS deployment is limited by the interference. Thus, it is necessary to position the FlyBS to maximize the SINR of the UEs served by the FlyBS, while interference to the BSs and other FlyBSs is minimized.

Therefore, in this section, we present a joint solution for the positioning of the FlyBSs and the association of the UEs, exploiting information related to the communication channel. We propose two novel algorithms for the joint positioning and association, considering the UEs' requirements on data rates. The first developed algorithm for the joint positioning and association is based on the PSO, while the second exploits the GA. Unlike other works, our objective is to maximize the UEs' satisfaction with the provided data rates. We show that the proposed joint positioning and association based on both PSO and GA notably outperforms a competitive state-of-the-art algorithm if the same amount of FlyBSs is deployed. We also discuss trade-offs between the PSO-based and GA-based solutions and assess their pros and cons.

This section is organized as follows. We start by defining system model and problem formulation.Then, the proposed algorithms for the association and the positioning are described, and implementation aspects are discussed. Followed by simulation scenario, a description of the competitive algorithm, and the performance evaluation.

4.4.1 System Model and Problem Formulation

In this section, we first define the system model for the positioning of the FlyBSs and for the association of the UEs. Then, we formulate the objective of this section.

System Model

We consider a set $\boldsymbol{N}$ of N UEs, where $n \in \boldsymbol{N}$ is a specific UE, a set $\boldsymbol{K}^S$ of K^S representing conventional SBSs, and a set $\boldsymbol{K}^F$ of K^F corresponding to the FlyBSs. Furthermore, we define a set of all BSs as $\boldsymbol{K} = \boldsymbol{K}^S \cup \boldsymbol{K}^F$with $K = K^S + K^F$ representing the total number of BSs. Note that the label "BS" represents both the SBSs and the FlyBSs in this section. The positions of the BSs are defined as $\boldsymbol{V} = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_K\}$, where $\boldsymbol{v}_k \in \mathbb{R}^3$, $k \in \boldsymbol{K}$ represents a position of the k-th BS. In the same way, we define a set of the

UEs' positions $\boldsymbol{U} = \{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_N\}$ with the position of the n-th UE denoted as $\boldsymbol{u}_n \in \mathbb{R}^3$, $n \in \boldsymbol{N}$. An activity status of the BS is indicated by a binary parameter, $\boldsymbol{\rho} = \{\rho_1, \ldots, \rho_K\}$. Setting $\rho_k = 1$ and $\rho_k = 0$ means that the k-th BS is being turned on and off, respectively.

The n-th UE and the k-th BS communicate over a radio channel, with SINR defined as:

$$\gamma_{n,k} = \frac{\rho_k P_k^{tx} |h_{n,k}\left(\boldsymbol{u}_n, \boldsymbol{v}_k\right)|^2}{\beta_{n,k}\sigma^2 + \sum_{l \in \boldsymbol{K}, l \neq k} \rho_l P_l^{tx} |h_{n,l}\left(\boldsymbol{u}_n, \boldsymbol{v}_l\right)|^2}, \tag{4.49}$$

where P_k^{tx} is the transmission power of the k-th BS, $h_{n,k}(\boldsymbol{u}_n, \boldsymbol{v}_k)$ is the channel realization between the k-th BS and the n-th UE, σ^2 is the noise power, and $\beta_{n,k} \in \boldsymbol{\beta} | \beta_{n,k} \in \{0, 1\}, \forall n \in \boldsymbol{N}, \forall k \in \boldsymbol{K}$ is the amount of bandwidth allocated for communication of the n-th UE with the k-th BS. The matrix $\boldsymbol{\beta} \in \boldsymbol{B}$ contains the bandwidth allocations of all BSs, with $\boldsymbol{B}$ representing a set of all feasible bandwidth allocations. Note that SINR is calculated only for active BSs (i.e., the BSs with $\rho_k = 1$). The bandwidth allocation also contains information about the UE's association to the BSs. The UE is considered to be associated to the BS to which it has non-zero $\beta_{n,k}$. We assume that the UE is associated to a single BS (i.e., the bandwidth for each UE is allocated at most to one BS). Then, the data rate provided by the k-th BS to the n-th UE via channel with the bandwidth B_k is defined as:

$$c_{n,k} = \beta_{n,k} B_k log_2\left(1 + \gamma_{n,k}\right) \tag{4.50}$$

Objective Formulation

Our objective is to find the positions of the FlyBSs and associate the UEs to the BSs in order to maximize the number of UEs satisfied with their experienced data rate. Without loss of generality, we focus on downlink direction. Note that the n-th UE is assumed to be satisfied if it experiences data rate $c_{n,k}$ equal to or higher than the minimum required data rate $c_n^{\min}$ (i.e., the UE is satisfied if $c_{n,k} \geq c_n^{\min}$). The BSs that are unused or cannot improve the UEs' satisfaction are turned off to save energy. Thus, our objective is to determine the optimal positions of the FlyBSs $\boldsymbol{V}^*$, the association of the UEs $\boldsymbol{\beta}^*$ (represented via bandwidth allocation), and the status of the BSs (on/off) $\boldsymbol{\rho}^*$. This objective is formulated as:

$$\boldsymbol{\beta}^*, \boldsymbol{V}^*, \boldsymbol{\rho}^* = \underset{\boldsymbol{\beta} \in \boldsymbol{B}, \boldsymbol{V} \in \mathbb{R}^{3 \times K}, \rho_k \in \{0,1\}}{\arg\max} \sum_{n \in \boldsymbol{N}} \sum_{k \in \boldsymbol{K}} \left[c_{n,k} \geq c_n^{\min}\right] \tag{4.51}$$

$$\text{subject to } \sum_{n \in \boldsymbol{N}} \beta_{n,k} \leq 1, \forall k \in \boldsymbol{K}, \tag{4.52}$$

$$\sum_{k \in \boldsymbol{K}} \left[\beta_{n,k} > 0\right] \leq 1, \forall n \in \boldsymbol{N}, \tag{4.53}$$

where the operator [.] is equal to 1 if the condition (e.g., $c_{n,k} \geq c_n^{\min}$) is fulfilled, otherwise it is equal to 0. The constraint (4.52) ensures that the BSs do not allocate more bandwidth than available. Furthermore, the constraint (4.53) ensures that each UE can be associated to a maximum of one BS.

4.4.2 Proposed Solution

The defined objective is an NP-hard problem (due to its definition as a non-convex function). Hence, to find the optimal positions of the FlyBSs, we exploit two evolutionary algorithms: GA and PSO [87]. The evolutionary algorithms iteratively search for the optimum within the search space, using several operations introduced later in this section.

First, we describe a general algorithm for the association of the UEs to the BSs, including a bandwidth allocation and a decision on the number of active BSs. Then, we integrate the association algorithm into the proposed algorithms for positioning of the FlyBSs based on the PSO and the GA, respectively. Last, a discussion of the practical implementation aspects is provided at the end of this section.

Association of the UEs and Bandwidth Allocation

The objective of the UEs' association is to determine the serving BS for each UE and to allocate bandwidth to the UEs to satisfy the UEs' required data rates c_n^{min} (i.e., each UE is allocated exactly with the bandwidth required to reach the c_n^{min}). Then, based on the association, we decide which BSs should be turned off, as those BSs do not improve the UEs' satisfaction. Note that we do not target a problem of minimization of the number of active BSs. Such a problem is not straightforward, and possible extension of our proposed algorithms is left for future research.

The proposed association of the UEs and the bandwidth allocation is described in Algorithm 1. In the initial phase, the active FlyBSs are randomly deployed within the area (line 1). The SINR between each BS and UE is calculated according to (4.49) from a path loss model, following the same approach as the authors in [88] (line 4).

Then, the n-th UE is temporarily associated to the serving BS s_n (i.e., to the BS with the highest experienced SINR to minimize the bandwidth required to satisfy the UE's c_n^{min} (line 6)). Based on the temporal association, a set of vectors $\boldsymbol{A} = \{\boldsymbol{A}_1, \boldsymbol{A}_k, \ldots, \boldsymbol{A}_K\}$ is created. Each vector $\boldsymbol{A}_k$ from $\boldsymbol{A}$ represents a list of the UEs associated to individual BSs. The list of UEs is created by adding the n-th UE to the vector $\boldsymbol{A}_{s_n}$ corresponding to the serving BS s_n (i.e., $\boldsymbol{A}_{s_n} \leftarrow \boldsymbol{A}_{s_n} \cup n$ (line 7)). Next, we create the set $\boldsymbol{K'}$ containing indices of the set $\boldsymbol{A}$ sorted according to the number of UEs served by each BS in descending order (line 9). For purposes of the bandwidth allocation in the next steps, a temporal set, $\boldsymbol{N'} \subseteq \boldsymbol{N}$, is created (line 10). Subsequently, $\boldsymbol{A}_k$ is emptied for each BS in $\boldsymbol{K}$ (line 11). Based on the ordered set $\boldsymbol{K'}$, the bandwidth for communication is allocated to the UEs to fulfill the UEs' data rate requirements. The BSs allocate the bandwidth until no bandwidth is left (lines 13 to 24).

The bandwidth is allocated according to the UEs' SINR in descending order (i.e., the UE with the highest SINR is allocated first). In terms of the algorithm, the UE n^* with the highest $\gamma_{n^*,k'}$ is selected first (line 15). Next, the bandwidth required to satisfy the UE n^* at the k-th BS is calculated as $\beta^{req} = c_{n^*}^{\min}/log_2\left(1+\gamma_{n^*,k'}\right)$ (line 16). If the k'-th BS has enough bandwidth, the β^{req} is allocated to the UE n^* and the UE n^* is associated to the k'-th BS (lines 17 to 19). The associated UE n^* is removed from the set $\boldsymbol{N'}$, and

the available bandwidth of the k'-th BS is updated (lines 20 and 21). The bandwidth allocation continues until there is not enough remaining bandwidth that can satisfy any UE. Then, the remaining bandwidth of each BS is divided among all the UEs served by the given BS (line 25). Finally, bandwidth allocation is divided by the B_k to obtain normalized bandwidth allocation $\beta_{n,k}$ (line 27), and the BSs serving no UE are turned off (line 28).

Algorithm 1 Association of UEs and bandwidth allocation.

1: Deploy FlyBSs by generating random positions $\boldsymbol{V}$; set $\rho_k = 1, \forall k \in \boldsymbol{K}$.
2: **for** $n \in \boldsymbol{N}$ **do**
3: **for** $k \in \boldsymbol{K}$ **do**
4: Calculate $\gamma_{n,k}$ via (4.49).
5: **end for**
6: $s_n \leftarrow \arg\max_{k \in \boldsymbol{K}} \gamma_{n,k}$
7: $\boldsymbol{A}_{s_n} \leftarrow \boldsymbol{A}_{s_n} \cup n$
8: **end for**
9: $\boldsymbol{K'} \leftarrow$ indices of $\boldsymbol{A}$ sorted in descending order.
10: $\boldsymbol{N'} \leftarrow \boldsymbol{N}$
11: $\boldsymbol{A}_k \leftarrow \emptyset, \forall k \in \boldsymbol{K}$
12: **for** $k' \in \boldsymbol{K'}$ **do**
13: **while** $B_{k'} > 0$ **do**
14: **for** $n \in \boldsymbol{N'}$ **do**
15: $n^* \leftarrow \arg\max_{n \in \boldsymbol{N'}} \gamma_{n,k'}$
16: $\beta^{req} = \frac{c_{n^*}^{\min}}{log_2\left(1+\gamma_{n^*,k'}\right)}$
17: **if** $B_{k'} \geq \beta^{req}$ **then**
18: $\beta_{n^*,k'} \leftarrow \beta^{req}$
19: $\boldsymbol{A}_{k'} \leftarrow \boldsymbol{A}_{k'} \cup n^*$
20: $\boldsymbol{N'} \leftarrow \boldsymbol{N'} \backslash n^*$
21: $B_{k'} \leftarrow B_{k'} - \beta^{req}$
22: **end if**
23: **end for**
24: **end while**
25: $\beta_{n',k} \leftarrow \beta_{n',k} + \frac{B_{k'}}{|\boldsymbol{A}_{k'}|}, \forall n \in \boldsymbol{A}_{k'}$
26: **end for**
27: $\beta_{n,k} \leftarrow \frac{\beta_{n',k}}{B_k}, \forall k \in K$
28: $\{\rho_k \leftarrow 0 | \forall k' \in \boldsymbol{K'}, \boldsymbol{A}_{k'} = \emptyset\}$

Positioning of FlyBSs via Particle Swarm Optimization

In this subsection, we describe the proposed algorithm for optimization of the FlyBSs' positions based on the PSO and its integration with the association. We exploit a common PSO described in [91] and adapt it to the objective defined in (4.51). The PSO searches for the optimal solution via a set of $l \in L$ particles $\{\boldsymbol{W}^1(t), \boldsymbol{W}^2(t), \ldots, \boldsymbol{W}^L(t)\}$ over iterations represented by the discrete time t. In our case, each particle contains the positions of all FlyBSs (i.e., $\boldsymbol{W}^l(0) = \boldsymbol{V}$). The search is done by updating the positions of the FlyBSs

via a velocity vector $\boldsymbol{D}^l(t)$ calculated as:

$$\begin{aligned}\boldsymbol{D}^l(t) = \phi \boldsymbol{D}^l(t-1) + c_p\phi_1 \left(\boldsymbol{W}^{l,local} - \boldsymbol{W}^l(t-1)\right) + \\ c_g\phi_2 \left(\boldsymbol{W}^{global} - \boldsymbol{W}^l(t-1)\right),\end{aligned} \tag{4.54}$$

where ϕ is the inertia weight determining the convergence speed, ϕ_1 and ϕ_2 are positive random variables, and c_p and c_g are the personal and global learning coefficients, respectively. The velocity vector represents a weighted sum of the previous velocity vector $\boldsymbol{D}^l(t-1)$, the difference between the FlyBSs' positions of the l-th particle $\boldsymbol{W}^l(t-1)$, and the l-th particle's local best solution $\boldsymbol{W}^{l,local}$ (i.e., historically the best FlyBSs' positions of the l-th particle), and the difference between the l-th particle $\boldsymbol{W}^l(t-1)$ and the global best solution $\boldsymbol{W}^{global}$. The global best solution $\boldsymbol{W}^{global}$ contains the best particle (i.e., the position of the FlyBSs with the highest targeted metric) out of all particles L, up to the current iteration t. In our objective, $\boldsymbol{D}^l(t)$ is a directional vector of the l-th particle, represented by the positions of all FlyBSs between the time instants t and $t-1$. Note that $\boldsymbol{D}^l(t)$ is calculated separately for each FlyBS of the l-th particle.

An example of the FlyBS position update is shown in Figure 4.17, where a single selected FlyBS at a position from the corresponding l-th particle $\boldsymbol{W}^l(t-1)$ is updated by $\boldsymbol{D}^l(t)$, considering the local and the global best positions of the selected FlyBS according to (4.54).

Each particle has its suitability represented by a cost function stored in Q^l. The suitability of the FlyBSs' positions and the UE's association of the l-th particle is defined by the cost function Q^l, reflecting our objective to maximize the UEs' satisfaction according to (4.51). Thus, the cost function is formulated as:

$$Q^l = \begin{cases} \sum_{n\in\boldsymbol{N}} \sum_{k\in\boldsymbol{K}} c_{n,k} & \text{if } c_{n,k} \geq c_n^{\min}, \forall n \in N \\ \sum_{n\in\boldsymbol{N}} \sum_{k\in\boldsymbol{K}} \left[c_{n,k} \geq c_n^{\min}\right] & \text{otherwise.} \end{cases} \tag{4.55}$$

The search for the optimal solution of the objective function is then achieved by updating the positions of the FlyBSs corresponding to each particle $\boldsymbol{W}^l(t)$ via a maximization of the particles' local best cost ($Q^{l,local}$) and a global best cost (Q^{global}). In other words,

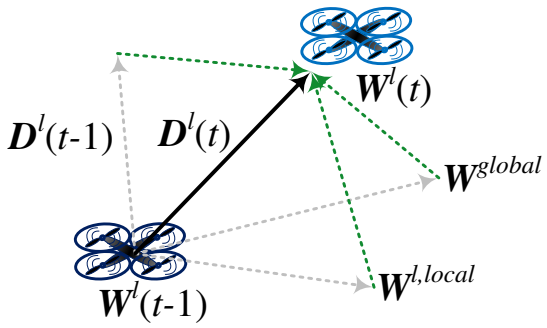


Figure 4.17. Update of FlyBS position via the proposed algorithm based on PSO.

the new position of the FlyBS is determined based on the best position of the given FlyBS represented by the l-th particle in the past, and the FlyBS's best position among all particles L. If all UEs are satisfied with the provided data rates, the remaining bandwidth is allocated to the UEs so that the sum of the UEs' data rates is maximized.

The proposed algorithm for the positioning based on the PSO with integration of the UEs' association is described in Algorithm 2. The PSO algorithm starts with the updated association with the unused BSs turned off (as explained in the previous subsection) (line 1). Based on the association, the particles are initialized (line 2). Then, the cost function of each particle is calculated via (4.55) (line 3). The particle with the highest cost is set as $\boldsymbol{W}^{global}$ and the cost of this particle is set to Q^{global} (lines 4 and 5). Then, the PSO iteratively updates the FlyBSs' positions until a maximum number of iterations M_{it} is reached (line 8). For each updated FlyBSs' position, the UEs' are re-associated (lines 9). Based on the updated FlyBSs' positions and the UEs' association, a suitability of the particle is evaluated via the cost function (4.55) (line 10). Then, we check if the updated positions improve the local solution (lines 11 to 13) or even the global solution (lines 15 to 17). Once all M_{it} iterations are completed, the $\boldsymbol{W}^{global}$ contains the set of FlyBSs' positions with the highest cost (suitability).

Algorithm 2 PSO for FlyBS positioning & UEs' association

1: Associate UEs & allocate bandwidth by Algorithm 1 with unused BSs turned off.
2: Initialize particles $\boldsymbol{W}^l(0)$, $l = 1, \ldots, L$ based on assoc.
3: $Q^{l,local} \leftarrow Q^l\left(\boldsymbol{W}^l(0)\right)$ via (4.55)
4: $Q^{global} \leftarrow \arg\max_{l \in L} Q^{l,local}$.
5: $\boldsymbol{W}^{global} \leftarrow \arg\max_{l \in L} \boldsymbol{W}^l(0)$.
6: **for** $t = 1, \ldots, M_{it}$ **do**
7: **for** $l = 1, \ldots, L$ **do**
8: $\boldsymbol{W}^l(t) = \boldsymbol{W}^l(t-1) + \boldsymbol{D}^l(t)$ via (4.54).
9: Assoc. UEs & alloc. bandwidth by Algorithm 1 with unused BSs turned off.
10: $Q^l(t) \leftarrow Q^l\left(\boldsymbol{W}^l(t)\right)$ via (4.55).
11: **if** $Q^l(t) > Q^{l,local}$ **then**
12: $Q^{l,local} \leftarrow Q^l(t)$
13: $\boldsymbol{W}^{l,local} \leftarrow \boldsymbol{W}^l(t)$
14: **end if**
15: **if** $Q^{l,local} > Q^{global}$ **then**
16: $Q^{global} \leftarrow Q^{l,local}$
17: $\boldsymbol{W}^{global} \leftarrow \boldsymbol{W}^{l,local}$
18: **end if**
19: **end for**
20: **end for**

Positioning of FlyBSs via Genetic Algorithm

In this subsection, we describe the proposed algorithm for optimization of the FlyBSs' positions based on the GA. We exploit a common GA described in [87] and adapt it to the optimization problem defined in (4.51). The GA consists of a population $\boldsymbol{G} =$

$\{\boldsymbol{g}^1, \ldots, \boldsymbol{g}^L\}$ with a size L. The population is composed of individuals $\boldsymbol{g}^l$ representing possible solutions (i.e., sets of the positions of the FlyBSs). Each individual consists of genes $\boldsymbol{g}_k^l$ corresponding to the positions of FlyBSs (i.e., $\boldsymbol{g}_k^l = \boldsymbol{v}_k$).

The first step of the GA is to generate an initial population with the size L. After that, a crossover operation inherent to all genetic algorithms is applied to the initial population. The crossover operation is understood as a mechanism during which new offspring are created from two selected parents. While each parent represents one of the previous positions of the given FlyBS, the new offspring defines a new possible position of the FlyBS. The selection of the parents j_1 and j_2 is done via Roulette Wheel Selection (RWS). The RWS selects parents based on their probability of survival defined by the cost function [153]. In our algorithm, the RWS is implemented by choosing the parent j via $\{j \in \mathbb{Z}, j \leq L | \sum_{l=1}^{l=j} F_l \geq \omega\}$ (i.e., by selecting a parent with fitness F_l equal to or larger than ω). The ω is selected randomly with uniform distribution $U(0,1)$, and F_l is determined from the fitness function defined as:

$$F_l = \frac{e^{-\frac{\overline{Q^l}}{max\{\overline{Q^l}\}}}}{\sum_{l \in L} e^{-\frac{\overline{Q^l}}{max\{\overline{Q^l}\}}}} \tag{4.56}$$

where $\overline{Q^l}$ is the normalized cost of the l-th individual calculated as $\overline{Q^l} = \frac{Q^l}{\sum_{l \in L} Q^l}$. Note that for the GA, we use the same cost function as expressed for the PSO in (4.55).

The number of offspring (new possible positions of the FlyBS) generated by the GA in each iteration is defined as $\lfloor Lp_c \rfloor$, where p_c is the crossover ratio representing a percentage of the whole population selected as the parents. The positions of the FlyBSs belonging to the selected parents are combined via an arithmetic recombination. This means the generated positions of the FlyBSs are influenced by a recombination parameter α denoting portions of the positions, which are taken from each of the selected parents. The parameter α is selected randomly from the uniform distribution $U(0,1)$, following an arithmetical crossover [153], where each offspring inherits a part of each parent's position.

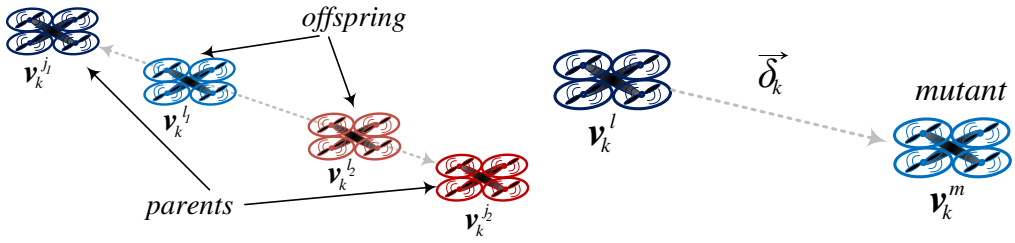
The principle of crossover operation is illustrated in Figure 4.18a, where two new offspring l_1 and l_2 (i.e., new possible positions of the FlyBS) are generated from the selected parents j_1 and j_2 (i.e., positions of the FlyBS in the past). The crossover operation takes positions of the parents $\boldsymbol{v}_k^{j_1}$ and $\boldsymbol{v}_k^{j_2}$ and modifies them as follows:

$$\boldsymbol{v}_k^{l_1} = \left(\alpha \boldsymbol{v}_k^{j_1} + (1-\alpha)\, \boldsymbol{v}_k^{j_2}\right) \tag{4.57}$$

$$\boldsymbol{v}_k^{l_2} = \left(\alpha \boldsymbol{v}_k^{j_2} + (1-\alpha)\, \boldsymbol{v}_k^{j_1}\right) \tag{4.58}$$

To preserve a diversity in the population, a mutation is exploited besides the crossover operation. The mutation corresponds to the process during which the position of the FlyBS ($\boldsymbol{v}_k^l$) is modified by a vector $\vec{\delta_k}$ as follows:

$$\boldsymbol{v}_k^m = \boldsymbol{v}_k^l + \vec{\delta_k}, \forall k \in \boldsymbol{K}^F \tag{4.59}$$



(a) Crossover. (b) Mutation.

Figure 4.18. Update of the FlyBS position by the proposed algorithm based on genetic algorithm.

The value of $\vec{\delta_k}$ is limited by the FlyBSs' deployment area (i.e., simulation area). Note that a direction of the vector $\vec{\delta_k}$ is selected randomly. In our case, the mutations are applied on randomly selected individuals with a probability p_m (i.e., there are $\lfloor L\,(L + Lp_c)\,p_m \rfloor$ mutants generated in each iteration). The principle of mutation is depicted in Figure 4.18b, where the l-th FlyBS with the position $\boldsymbol{v}_k^l$ generates a mutant m with a new position corresponding to $\boldsymbol{v}_k^m$.

After the crossovers and mutations, whole population (i.e., parents, offspring, and mutants) is evaluated via the fitness function (4.56) to select proper individuals (i.e., sets of positions of the FlyBSs) for the new iteration.

The proposed GA solution exploiting the defined operations (crossover, mutation, etc.) is described in Algorithm 3. The algorithm is initialized by updated association with unused BSs turned off by Algorithm 1 (line 1). The population $\boldsymbol{G}$ is then generated based on the association (line 2) with the cost of each individual determined according to (4.55) (line 3). The crossovers and mutations are applied to the positions of FlyBSs in the $\boldsymbol{G}$ (lines 6 and 7) to generate new possible solutions (i.e., sets of positions of the FlyBSs) within the constrained area of search for the optimum. Due to the updated positions of the FlyBSs, the UEs' association and the bandwidth allocation are updated as well. The UEs are associated by considering only the BSs that are turned on (line 8). The cost of the updated population is calculated via (4.55) (line 9). Based on the fitness function (4.56), the fittest individuals are selected for the next iteration (line 11). If the population is not diverse (i.e., the cost of all individuals is the same), the mutation percentage p_m is increased to $p_{mutate,high}$ to avoid premature convergence to a non-optimum solution; otherwise, p_m is kept at p_{mutate} (line 12). Then, the individual with the highest cost Q^l (suitability) is selected as the most suitable solution (line 13).

Practical implementation aspects

In this subsection, we discuss aspects related to implementation of the UEs' association and the positioning of the FlyBSs. First, we focus on the mobile network entities where the

Algorithm 3 GA for FlyBS positioning & UEs' association

```
Assoc. UEs & allocate bandwidth by Algorithm 1 with unused BSs turned off
Initialize population G based on association
Calculate cost Q^l of g^l, ∀l ∈ L via (4.55)
for t = 1, ..., M_it do
    for l = 1, ..., L do
        Apply crossovers via (4.57) and (4.58).
        Apply mutations via (4.59).
        Assoc. UEs & alloc. bandwidth by Algorithm 1 with unused BSs turned off
        Calculate cost Q^l of g^l, ∀l ∈ L via (4.55)
    end for
    Select the fittest individuals to next iteration via (4.56)
    Check population diversity and adjust p_mutate.
    g* ← arg max_{l∈L} cost(g^l)
end for
```

algorithms for the joint positioning and association can be deployed and run. The most straightforward option is to implement the algorithm directly at the FlyBSs. On one hand, this option leads to a low latency of determining FlyBSs' positions and UEs' association, since the FlyBSs just exchange control information among themselves, and there is no need to communicate with the core network. On the other hand, this solution also drains batteries of the FlyBSs; thus, the operational time of the FlyBSs is reduced. Although the common UAVs, such as quad- or hexa-copters, can fly several hours if they are powered with hydrogen cells, any additional energy consumption is undesirable [10]. Thus, running the proposed algorithm directly at the FlyBSs is limited only to the scenarios where a short operation time of the FlyBSs is not a problem.

The second option is to run the algorithm in a fixed infrastructure, such as common SBSs, core network, or a Base Band Unit (BBU) if the Cloud-Radio Access Network (C-RAN)) is deployed [10]. In this case, the energy required to run the proposed algorithm for the positioning and association is not that critical. However, the latency (especially if the algorithm is run in the BBU connected through a non-ideal fronthaul [154]) is the main concern here and can result in incorrect positioning of the FlyBSs. As a consequence, this option is preferable if a higher latency does not degrade the performance of the proposed algorithm, such as for slow-moving UEs (pedestrians), where the delay on the order of tens of milliseconds plays no role due to the slow movement of the UEs.

The proposed algorithms require information about the UEs' positions (these are assumed to be known in, e.g., [73] or [75]), their required data rates, and environment for estimation of the propagation losses (such as [155] or [156]) to determine the SINR for a given FlyBS's positions (as assumed, e.g., in [88] and outlined in [73]). The required data rate is known to the network, as this information is required for scheduling. The UEs' positions represent overhead on the order of tens of bytes, which is negligible.

4.4.3 Simulation Scenario and Performance Evaluation

Performance of the proposed solution is analyzed and compared with a competitive solution by simulations conducted in MATLAB.

Simulation Scenario

We assume a scenario in which the deployment of the FlyBSs is meaningful (i.e., the UEs benefit from increased data rate satisfaction). Thus, four small-cell SBSs (i.e., $K^S = 4$), with transmission power of 15 dBm, are deployed at positions [400 400, 400 1200, 1200 400, 1200 1200] in a simulation area of 1600 m x 1600 m, as shown in Figure 4.19. Moreover, up to twenty FlyBSs (i.e., $K^F = 20$), with the same transmission power as the SBSs, are deployed in the same area. The deployment of both the FlyBSs and the SBSs emulates a realistic case in which the FlyBSs cooperate with existing infrastructure, and interference among the FlyBSs and the SBSs plays an important role in the association and positioning. Thus, we also assume that all BSs transmit on the same frequency (i.e., each BS interferes with other BSs). A signal propagation for the SBSs is modeled according to [157] with path loss model $PL = 128.1 + 37.6log_{10}d$, where d is a distance between the UE and the SBS. For the FlyBSs, we select a commonly used path loss model from [155], with Suburban environment parameters from [158]. A connectivity of the FlyBSs to the core network of the operator is assumed to be of a sufficient capacity to transfer all the UEs' data transmitted over the access link (from the UE to the FlyBS) as expected (e.g., in [10] and [159]). The major parameters of the simulations are summarized in Table 4.2.

Performance Evaluation

In this section, we provide a performance evaluation of the proposed solutions. The performance of the proposed algorithms based on GA and PSO is compared with a commonly exploited k-means algorithm (see, e.g., [93]) extended with the bandwidth allocation according to our proposed Algorithm 1 for a fair comparison. To the best of our

Table 4.2. Simulation parameters.

Parameter	**Value**
Simulation area	1600m x 1600m
Carrier frequency	2 GHz
Number of SBSs/Maximal number of FlyBSs	4/20
Tx power of SBS/FlyBS	15/15 dBm
Bandwidth of SBS/FlyBS	20/20 MHz
SBS/FlyBS/UE height	20/20/1.5 m
Maximal number of iterations GA/PSO/k-means	100/100/100
Population size GA/Number of particles PSO	100/100
GA - $p_c/p_{mutate}/p_{mutate,high}$	0.8/0.3/0.8
PSO - $\phi/\phi_1/\phi_2/c_g/c_p$	4.1/2.05/2.05/1.5/1.5
Number of simulation drops	1000 drops

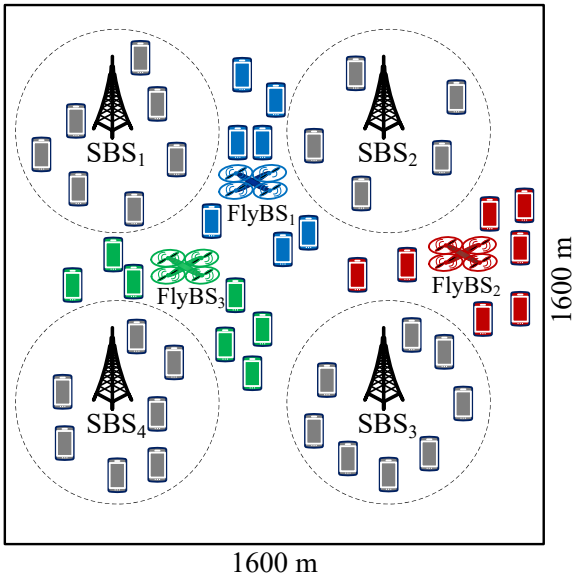
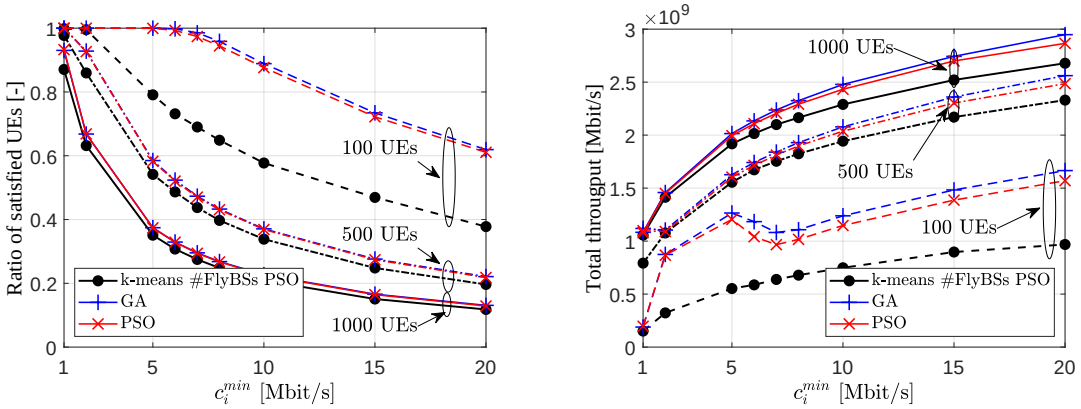


Figure 4.19. Simulation area with FlyBSs and SBSs and associated UEs (association to individual FlyBSs is indicated by colors).

knowledge, there is no other algorithm for comparison that solves joint association and positioning and targets maximization of the UEs' satisfaction. As exploiting the k-means requires that we know the number of FlyBSs to be deployed, we investigate the k-means with the same number of FlyBSs as the number of FlyBSs required by our proposed algorithm based on the PSO. This allows a fair comparison of the k-means and the PSO in terms of the UE's satisfaction, which is our major objective.

In Figure 4.20a, we show the ratio of the satisfied UEs to the achieved throughput (i.e., the UEs for which $c_{n,k} \geq c_n^{\min}$) for $c_n^{\min}$ set to the same value for all the UEs and ranging between 1 and 20 Mbit/s. For all compared algorithms, the ratio of satisfied UEs is decreasing with increase of both $c_n^{\min}$ and the number of UEs. The decrease in satisfaction is because, while the UEs require higher data rates (i.e., higher $c_n^{\min}$), the BSs still have a limited amount of bandwidth that can be allocated to the UEs. The gain of the proposed algorithms based on GA and PSO compared to the k-means and 100 UEs is up to 30 % and 31%, respectively. Although the gain decreases with more users in the area, the gain of the proposed algorithms is above 6% for almost all values of $c_n^{\min}$, even for 1000 UEs. The improvement in the UEs' satisfaction is achieved by the positioning of the FlyBSs, association of the UEs, and turning off the BSs that are not necessary. All these aspects lead to a higher level of the received signal and/or lower interference from neighboring BSs.

Figure 4.20b depicts a total throughput, which is defined as the sum of data rates $c_{n,k}$ over all UEs. It is demonstrated that the total throughput is increased by the proposed algorithms (GA and PSO) when compared to the k-means. The gain typically ranges between 19% and 47%. Again, the highest gain is achieved for a lower number of UEs. For example, for 100 UEs, both proposed algorithms lead to an improvement in the total throughput by approximately 30% for $c_n^{\min}$= 1 Mbit/s with respect to the k-means. The



(a) Ratio of the UEs satisfied with experienced throughput.

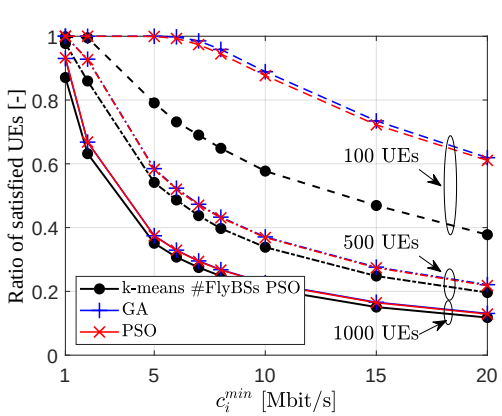
(b) Total throughput of all UEs in the network.

Figure 4.20. Improvement of network performance in (a) UEs satisfaction and (b) network throughput.

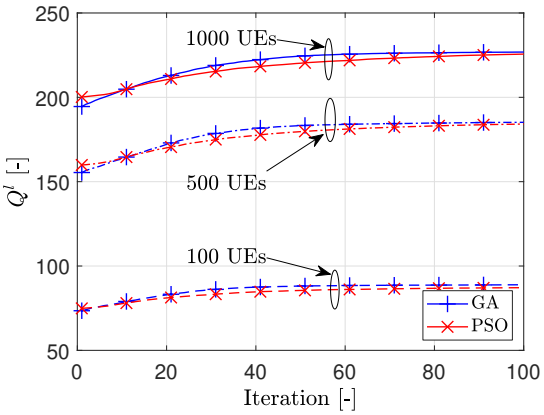
highest gain is observed for $c_n^{\min} = 2$ Mbit/s and 100 UEs; in this case, the proposed algorithms based on GA and PSO outperform the k-means by 128% and 126%, respectively. This notable gain is because the k-means fails to handle small groups of the UEs requiring relatively high data rates. In contrast, both GA- and PSO-based algorithms converge to a suitable deployment and association that avoids redundant interference.

Figure 4.20b also shows an interesting phenomenon, as the total throughput for 500 UEs is higher than that for 1000 UEs if $c_n^{\min} = 1$ Mbit/s. This is due to the definition of the cost function in (4.55), and the fact that the primary objective is to maximize the UEs' satisfaction while the throughput maximization is only a secondary objective (see (4.55)). Consequently, if all UEs' requirements are satisfied with the experienced throughput (i.e., the case with 500 UEs and $c_n^{\min} = 1$ Mbit/s, as shown in Figure 4.20a), both the GA and the PSO start maximizing the total throughput as well. In contrast, if some of the UEs remain unsatisfied (i.e., the case with 1000 UEs deployed in the area and $c_n^{\min} = 1$ Mbit/s), the GA and the PSO aim to maximize the UEs' satisfaction while the throughput maximization does not take place. Consequently, the total throughput for 1000 UEs is slightly lower than the total throughput achieved for 500 UEs if $c_n^{\min} = 1$ Mbit/s.

The same behavior can be seen also for 100 UEs, where the throughput gradually increases as long as $c_n^{\min} \leq 5$ Mbit/s since all 100 UEs are always satisfied (i.e., the proposed algorithms are able to further maximize the throughput according to the secondary objective). However, for $c_n^{\min}$ higher than 5 Mbit/s, the total throughput starts decreasing (see the throughput for $c_n^{\min}$ between 5 and 7 Mbit/s). For these values of $c_n^{\min}$, the case when all UEs are satisfied (i.e., when throughput maximization takes place) occurs for some simulation drops while, in other drops, not all UEs are satisfied (i.e., the secondary objective of throughput maximization does not take place). The cases when all UEs are satisfied and when some UEs are not satisfied are represented by different slopes of the



(a) Number of deployed flying base stations in order to reach the UEs' satisfaction presented in Figure 4.20a.



(b) Evolution of cost over iterations of the GA and PSO for $c_n^{\min} = 10$ Mbit/s.

Figure 4.21. Improvement of network performance in (a) UEs satisfaction and (b) network throughput.

total throughput over $c_n^{\min}$. Combination of both cases for $c_n^{\min}$ between 5 and 7 Mbit/s results in a decrease in the total throughput. Nonetheless, with a further increase in $c_n^{\min}$ above 7 Mbit/s, the total throughput starts increasing again, following the slope corresponding to the second case (some UEs are not satisfied), with a slower increase in the total throughput comparing to $c_n^{\min} \leq 5$ Mbit/s. The slope of the second case ($c_n^{\min}$ ¿ 7 Mbit/s) is lower because the secondary objective takes place in no (or almost no) drops. Note that the first non-satisfied UEs are those with the worst channel quality. Thus, the total throughput still increases even if the secondary objective is not considered. This is due to the allocation of bandwidth to the UEs with a higher channel quality.

The number of active FlyBSs required to maximize the UEs' satisfaction is presented in Figure 4.21a. The number of FlyBSs for the k-means is not shown, as the k-means cannot change the number of active FlyBSs, and we set it to the number of active FlyBSs required by the PSO, as explained earlier. From Figure 4.21a, we can see that the number of active FlyBSs with required throughput increases for low c_n^{min}, but for c_n^{min} above 5 Mbit/s, the number of active FlyBSs starts slowly decreasing for 500 and 1000 UEs. The decrease in the number of active FlyBSs for a larger c_n^{min} is due to the fact that the additional FlyBSs increase interference more than the amount the UEs can gain from the improved level of the useful signal provided by the serving BS.

The proposed algorithms based on GA and PSO find the solution iteratively. An evolution of the cost function Q^l with iterations for $c_n^{\min} = 10$ Mbit/s is shown in Figure 4.21b. The figure depicts the number of UEs satisfied with their data rates in each iteration (following (4.55)). The value of the cost function (i.e., the number of satisfied UEs) iteratively improves as the algorithms based on both the GA and the PSO find better positions of the FlyBSs. It is shown that the positioning based on the GA provides slightly higher values of the cost function in comparison to the PSO (difference is on the

order of a few percent. However, the cost function converges almost to its maximum in roughly 30 or 40 iterations for both algorithms. Then, the cost function remains almost constant. Such a low number of iterations required for the convergence makes the proposed solutions promising for real networks.

The performance analysis presented in the previous figures shows that the GA slightly improves the UEs' satisfaction with respect to the PSO (by up to 2%, see Figure 4.20a). At the same time, the GA increases the total throughput by 1%˜10% compared to the PSO (see Figure 4.20b). In addition, the GA converges to these gains while requiring approximately one FlyBS less than the PSO (see Figure 4.21a), so the operational cost is slightly reduced by the GA as well. However, this gain is at the cost of a higher time complexity of the algorithm based on the GA. The GA is of a higher time complexity than the PSO because of the nature of the base GA and PSO algorithms (see, for example, [160]). We express the time complexity as the time required to complete one iteration of the algorithm (i.e., one update of the FlyBSs' positions for each population/particle). The iteration takes 0.22 and 0.13 seconds, on average, for the GA and the PSO, respectively. Note that the times of each iteration are obtained at a desktop PC with Intel i7-7700K@4.2 GHz CPU and 32 GBs of RAM.

4.4.4 Conclusion

In this section, we have proposed an algorithm for the joint positioning of the FlyBSs and association of the UEs to maximize the number of UEs satisfied with the experienced data rates. The developed algorithm is presented in two variants: one based on the GA and one on the PSO. We show that both approaches improve the UEs' satisfaction compared to the commonly used k-means by up to roughly 30%. Also, a gain in the total throughput of all UEs is observed for both proposed algorithms. The gain typically varies between 19% to 47%, but reaches its maximum of more than 100% for scenarios with a lower number of UEs and medium to high data rates. We also show that the GA slightly increases the UEs' satisfaction and the total throughput while reducing the number of required FlyBSs compared to the PSO. This improvement is, however, at the cost of a higher time complexity.

The proposed algorithms exploit the collected network information from the UEs, as described in Section 4.1 and Section 4.2. The mobile network optimization, as presented in this section, is then applicable in general scenarios.

Chapter 5

Resource allocation the MEC

The MEC provides computation resources at the edge of the mobile network, that can be exploited by the mobile operator for the mobile network optimization, but also, by the mobile users for offloading of their computation intensive tasks from their UEs to the BSs serving as the MEC host. The previous chapter have dealt with collecting of the mobile network information from the UEs and optimization of the mobile network via deployment of the FlyBSs. In this chapter, we present a work for optimization of the mobile network, where the UEs exploit the MEC for the computation offloading. The optimization of the mobile network is done via determination of communication and computation resource allocation for the mobile UEs. The communication resources are represented by the communication bandwidth (in LTE-A represented by the RBs), while the computation resources by the CPU processing time and RAM. Due to security concerns, the computation resources are encapsulated and form either a VM or a container.

In general computing, including exploitation of the MEC the users require a certain level of service, represented by the QoS. One of the primary indicators of the QoS in the MEC is the offloading delay. The offloading delay represents time (or duration) from the time the offloading starts to collection of the results back at the UE. The offloading delay consists of: i) t_{O} the time required to deliver the offloaded task from the UE to the BS that starts the computation, ii) t_{P} the time required to process the offloaded task, iii) t_{C} the time required to deliver the processed data from the BS that finishes the computation to the UE, iv) t_{H} the time consumed by the handover process, v) t_{M} the time of the VM starts (including obtaining UE's application which processes offloaded tasks) during the offloading. The total offloading delay experienced by the UE is then $t_{\mathrm{MEC}} = t_{\mathrm{O}} + t_{\mathrm{P}} + t_{\mathrm{C}} + t_{\mathrm{M}} + t_{\mathrm{H}}$.

The rest of this chapter is organized as follows, first, a general description of MDP is provided, as it is the main mathematical tool, that is exploited for solving the communication and computation resource allocation in this chapter. Then, in Section 5.2 we present a solution for selection of the communication path for the computation offloading. In Section 5.3 we describe a joint communication and computation resource allocation with fixed mobility prediction accuracy. Then, this work is extended to a general scenario, where we propose a mobility and channel quality prediction framework and joint algorithms for

communication and computation resource allocation.

5.1 Markov Decision Process

The proposed algorithm for determination of the communication path, i.e., path selection, and computation resource allocation are based on the MDP. Therefore, we describe the MDP and how it is exploited in the proposed algorithms in this section.

The MDP is a discrete-time stochastic control process that provides a mathematical framework for modeling decision making. The MDP is an extension of Markov chains, differing in addition of actions and rewards. A MDP is a tuple $(S, A, P_a(s, s'), R_a(s, s'))$, where s is a set of states called state space, A is a set of actions called action space, $P_a(s, s')$ is a probability that action a in state s at discrete time k will lead to state s' at discrete time $k+1$ and $R_a(s, s')$ is the immediate reward received after transitioning from state s to state s' due to action a. The goal of solving the MDP is to find a policy denoted as π that maximizes total reward V_π^k over a potentially infinite horizon given by the discrete time k [161]:

$$V_\pi^k = Est(\sum_k R^t|\pi, s) = R(s) + \sum_k T(s, \pi(s,k), s')V_\pi^{k-1}(s') \tag{5.1}$$
$$\pi : s \to a,$$

where summation of reward per time step $(T(s, \pi(s,k), s')V_\pi^{k-1}(s'))$, represents expected future payoff as a sum over k steps, $Est()$ represents an estimate of reward. The estimation is necessary in the mobile networks, as the channel quality fluctuates frequently.

Lets demonstrate our MDP approach on an example of a chain of two consecutive time steps as shown in Figure 5.1, where states s (i.e., 1, 2, 3, and 4) represent a selected serving BS, that is exploited to transmit the offloaded task to the MEC host. At the beginning of the process (at time $t = 0$), we are in the state $s = 1$, as current serving BS is 1. In the next time step (i.e., $t = 1$), a transition to four different states is possible. It means the serving BS may stay the same, i.e., $s = 1$ or may change to a different one, i.e., $s' \in \{2, 3, 4\}$. These states are connected with the current state via edges denoting the immediate transition reward $R(s)$. The immediate transition reward of staying in the current state (s) is 0 as the serving BS remains the same (there is no handover) and thus no gain (loss) is obtained by this transition. In case of the transition to another state, i.e., s', there is a negative immediate reward due to handover (in Figure 5.1 denoted as $-t_H$). The handover introduces negative immediate reward as it leads to an overhead and no useful data are being transmitted during the handover. Nevertheless, in case of the transition to another state $(2, 3, 4)$, there is also expected future reward (in Figure 5.1 denoted as $T(s, \pi(s,k), s')$), which is calculated as a reward introduced by connection to a new BS (2, 3, or 4) and staying there for a duration of a one time step.

To obtain the total reward V_k^π, the procedure from Figure 5.1 has to be repeated multiple times (for all time steps) until all required data is transmitted as shown in

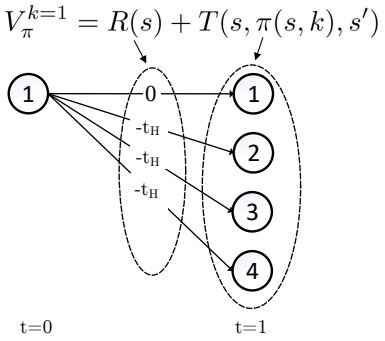


Figure 5.1. Reward on a chain of a single time step (i.e., $k = 1$).

Figure 5.2. Then, the selected path is represented by a chain of serving BSs (e.g., red line starting in state 1 in time $t = 0$ and ending in state 2 in time $t = k$).

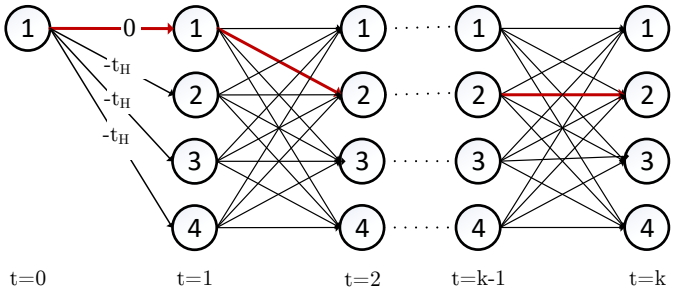


Figure 5.2. Chain for calculation of the total reward (for the sake of figure clarity, negative rewards $-t_H$ due to handover between states is not depicted for $t > 1$).

5.1.1 MDP in communication and computation resource allocation

The MDP is exploited for determination of communication path selection, as well as, for computation resource allocation. Therefore, an example how the MDP is mapped to the path selection problem is described. In the communication path selection problem, the immediate reward $R(s)$ is represented by handover delay $t_{\mathrm{H}}(q_s, q'_{s'})$ and energy consumed during handover $E_H(q_s, q'_{s'})$. The reward in terms of delay per step, in (5.1) denoted as $T(s, \pi(s,k), s')V_\pi^{k-1}(s'))$, is calculated as a difference in communication delay if path $q'_{s'}$ is selected instead of q_s for both radio (i.e., $t_{\mathrm{R}}(q_s) - t_{\mathrm{R}}(q'_{s'})$) and backhaul (i.e., $t_{\mathrm{B}}(q_s) - t_{\mathrm{B}}(q'_{s'})$). The reward in terms of UE's energy consumption per step is calculated as a difference in energy consumed if the communication over radio takes place via path $q'_{s'}$ instead of q_s(i.e., $E_R(q_s) - E_R(q'_{s'})$).

All rewards (delay and energy) are added up to obtain the total reward. We also reflect the possibility of weighting energy and delay (as defined in (5.3)) by parameter γ.

Thus, the total reward for our proposal is defined as:

$$V_{\pi}^{k}(q_s, q'_{s'}) = \gamma[-E_H(q_s, q'_{s'}) + \sum_{k}(E_R(q_s) - E_R(q'_{s'}))] + \\ (1-\gamma)[-t_{\mathrm{H}}(q_s, q'_{s'}) + \sum_{k}(t_{\mathrm{R}}(q_s) - t_{\mathrm{R}}(q'_{s'})) + \sum_{k}(t_{\mathrm{B}}(q_s) - t_{\mathrm{B}}(q'_{s'}))] \tag{5.2}$$

where $E_R(q_s)$ and $E_R(q'_{s'})$ denote the energy consumed by the UE's radio communication using current path q_s and the new path $q'_{s'}$, respectively, $t_{\mathrm{H}}(q_s, q'_{s'})$ and $E_H(q_s, q'_{s'})$ stands for the delay and the energy consumed by handover from the serving cell to the neighboring cell (transition from the path q_s to the path $q'_{s'}$), respectively. Delay due to the handover ($t_{\mathrm{H}}(q_s, q'_{s'})$) and energy consumed by the UE during handover procedure ($E_H(q_s, q'_{s'})$) reflect an overhead in terms of additional delay and energy consumption caused by the handover procedure, respectively.

5.2 Selection of communication paths for MEC

In this section, we describe an algorithm for selection of communication path exploiting handover in order to avoid distribution of the offloaded data via a backhaul of a limited capacity. Our motivation is to shorten the time necessary for transferring the offloaded data to an individual computing SCeNBs. To prevent high energy consumption at the UE side, the energy spent by the UE for data transmission as well as energy spent by handover itself is also considered for selection of the most suitable way of data delivery. The problem is formulated as a MDP. In the MDP, any change of the serving SCeNB (i.e., each handover), motivated by an improvement of data delivery, is rewarded depending on its impact on the UE's energy consumption and transmission delay caused by both data transmission and handover. The algorithm selects also the path for delivery of computation results back to the UE. Independent selection of the communication paths for data offloading (uplink) and results delivery to the UE (downlink) solves the problem of mobility management for users exploiting the MEC (i.e., the problem of users moving from one cell to another when the offloaded application is currently computed). Consequently, the MEC can be efficiently utilized also by the moving UEs. In addition, the algorithm is suitable for parallel computing, so, parallelized parts of the code (offloaded data) can be delivered to multiple computing cells via multiple routes to minimize the transmission delay.

This section is an extension of our previous work presented in [162], where we have proposed general framework for the path selection and we have provided basic performance analysis. With respect to [162], we extend our work in the following aspects: 1) we consider path selection not only for uplink data offloading but also for downlink reception of computation results in order to address a problem of user mobility management; 2) we present more detailed description of the proposed algorithm including implementation

aspects related to derivation of required parameters; 3) we enhance simulations by consideration of multi-user multi-cell scenario and user's mobility; 4) we evaluate also impact of the proposed algorithm on the load of the backhaul network.

The rest of this section is organized as follows. In the next section, the proposed algorithm for path selection is described along with implementation aspects. Simulation methodology and scenario are presented in Section 5.2.7. Section 5.2.8 provides performance evaluation and discussion of simulation results. The last section summarizes major conclusions and outlines potential future research work.

5.2.1 Path Selection Algorithm

In this section, we present system model and proposed algorithm for data delivery from the UE to individual BSs or SCeNBs performing computation and delivery of the computing results back to the UE. Designed path selection algorithm takes into account the UE's energy consumption (both energy spent by data transmission and handover), handover delay, radio channel quality, and backhaul conditions. Furthermore, we discuss the implementation aspects and a possible reduction of the computation complexity.

5.2.2 System model

We assume the system composed of S SCeNBs that act as MEC hosts and U UEs. Furthermore, for each UE, we define set X consisting of n computing SCeNBs and set I consisting of m SCeNBs that are in the neighborhood of the UE and can communicate with the UE directly through the radio link. Note that sets X and I may be fully or partially overlapping if computing cells are in radio communication range of the UE (i.e., if the UE can connect directly to the computing cells via radio link). In our model, the SCeNB providing the highest RSSI to the UE is selected to be the serving cell for each UE [109, 163]. In case of the UE's movement, the serving cell is updated if the RSSI from the target SCeNB ($RSSI_{TC}$) becomes higher than the RSSI of the serving cell ($RSSI_{SC}$) plus handover hysteresis ($\triangle_{HM}$), i.e., if $RSSI_{TC} > RSSI_{SC} + \triangle_{HM}$.

An example of the network model is shown in Figure 5.3. In this figure, c represents capacity of the link and upper indexes B and R stand for backhaul and radio links, respectively. In the given example, the cluster of cells performing computation is formed of four SCeNBs. Out of those SCeNBs, the $SCeNB_1$ is selected as the serving cell.

As depicted in Figure 5.3, data from the UE can be transferred to the $SCeNB_i$ over the radio link with capacity c_i^R. The $SCeNB_i$ is connected to the operator's core via the backhaul with capacity c_i^B. The offloaded data is processed by the $SCeNB_i$ or forwarded to another computing $SCeNB_x$ through backhaul of the $SCeNB_i$ (with capacity c_i^B in uplink) and backhaul of the computing cell $SCeNB_x$ (with capacity c_x^B in downlink). Note that index x stands for any SCeNB out of X except the $SCeNB_i$ (i.e., $x = 2, 3, 4$ in Figure 5.3). Selection of the computing cells can be done according to complexity of the offloaded data processing and available computing power of the SCeNBs as suggested, e.g., in [99, 121, 164]. After the SCeNBs finish data computation, the results are delivered back

to the UE. New path for backward delivery of computation results (from each $SCeNB_x$ to the UE) must be derived if radio and backhaul links are not symmetric in uplink and downlink, if the UE moves during computation, or if the channel/link load or quality changes. Therefore, computation results can be delivered to the UE through a new cell(s), which again minimize delay and/or energy consumption according to the user's preference.

In case when the offloaded task requires simultaneous uploading (task offloading) and downloading (results reception), both are done via the same serving cell selected by the proposed path selection algorithm. It means, the UE communicates over the same serving cell for uplink and downlink until all data is transmitted and received. Then, if handover is beneficial, new serving cell is selected (still just one serving cell for both uplink and downlink). Changing the UE's serving cell for offloading and results reception can lead to a change in assigned IP address and, thus, it may lead to problem with routing of data to destination. Nevertheless, this problem is solved by a method for addressing devices in MEC as outlined in [101].

Each computing task offloaded to the SCC (in this thesis denoted as "offloaded task") is divided between the computing cells. Each $SCeNB_x$ (where $x \in X$) is expected to compute a part $\lambda_x \in (0, 1]$ of the whole offloaded task, which is of the overall size of L_{UE}. The individual part L_x computed by the $SCeNB_x$ is then expressed as $L_x = \lambda_x \cdot L_{UE}$ where $\sum \lambda_x = 1$. In this section, we assume to split the offloaded task into parts with the same size among all computing cells, i.e., $\lambda_1 = \lambda_2 = ... = \lambda_x$. In general, the size of each offloaded part should correspond to the computing power of individual SCeNB involved in computation. The optimal distribution of offloaded task to individual computing SCeNBs is out of scope of this thesis and this topic is left for future research.

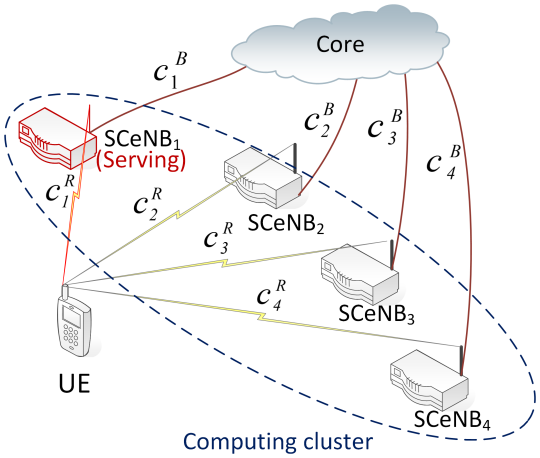


Figure 5.3. Network topology and definition of parameters required for path selection.

The common approaches for delivery of the offloaded task to the computing cells and back assume that data from the UE to the computing cells is always delivered through the same serving cell ($SCeNB_1$ in Figure 5.3) [19, 21]. This serving cell is selected only according to the rules applied in common mobile networks, i.e., with respect to radio channel quality or available capacity of radio channel [110, 163]. To overcome potential delay due to distribution of data among all computing $SCeNB_x$ over the backhaul with

limited throughput, we exploit an opportunity to transfer data also via neighboring cells. In this case, individual parts of the data for computation are delivered to individual computing cells through specific neighbors, which offer the lowest transmission delay over both radio and backhaul links.

Note that for each computing cell, data can be delivered through different neighboring cell. This implies a need for performing handover during communication. In our proposed algorithm, handover is not enforced during transmission of the offloaded task to each of assigned computing cells. Instead, handover is performed when no data is being transmitted at the moment. For example, following Figure 5.3, data designated to be computed at the $SCeNB_1$ is transmitted to this SCeNB. After successful transmission, handover is performed before data to the next SCeNB (e.g., $SCeNB_2$) is transmitted. Therefore, offloading is not interrupted by handover as each part of the task is offloaded/received at its destination before handover.

Comparing to the conventional handover in mobile networks, the handover is not conditioned only by radio quality but also by backhaul [111] and selection of computing cells. The proposed algorithm is labeled as a PSwH. The scheme using single serving cell selected in conventional way according to radio quality for delivery of all offloaded data is denoted as a SO in the rest of the thesis.

5.2.3 Path selection exploiting handover

The proposed path selection algorithm suitable for the SCC combines the time required for transmission of offloaded task over radio and backhaul links (in the rest of thesis denoted as *transmission delay*) and energy consumed by the UE by both transmission of data over radio link and handover (in the rest of thesis denoted as *energy*).

Each q-th path between the UE and the computing cell is described by the transmission delay (t_q) and the energy consumed by the UE's transmission (E_q). In our path selection algorithm, both t_q and E_q are combined into a single metric of the q-th path, M_q:

$$M_q = \gamma\overline{E_q} + (1-\gamma)\overline{t_q} \tag{5.3}$$

where γ is the weighting factor showing preference for low delay ($\gamma \to 0$) or for high energy efficiency ($\gamma \to 1$), $\overline{t_q}$ ($\overline{E_q}$) represents normalized delay (energy) of the q-th path. In order to enable combination of both metrics, both are normalized (i.e., scaled into range from 0 to 1) with respect to the maximum observed value as follows:

$$\overline{t_q} = \frac{t_q}{max\{t_1, t_2, ...t_p\}} \tag{5.4}$$

$$\overline{E_q} = \frac{E_q}{max\{E_1, E_2, ...E_p\}} \tag{5.5}$$

where p is the number of possible paths from the UE to the computing cells. Parameter p is calculated as the cardinality of Q, i.e., $p = |Q|$, where Q is the set of all possible

paths including all combinations of computing SCeNBs (set X) and all SCeNBs within radio communication range of the UE (set I). The path selection algorithm is defined as the MDP, as described in Section 5.1, where the state s represents currently selected path q_s (using the serving cell selected in conventional way according to radio link quality) and the future state s' is another possible path $q'_{s'}$ (composed of radio and backhaul connections) out of Q. Note that Q includes also paths obtained by performing handover to neighboring cells. As the radio and backhaul parameters fluctuate over time, calculated time of transmission is only an estimation of the expected transmission time. This estimation introduces an error in derivation of the reward. The estimation error can be avoided by reservation of radio and backhaul resources solely for the purposes of data offloading to the SCC, i.e., using Guaranteed Bit Rate (GBR) [165] in LTE-A networks. Nevertheless, this would lead to QoS degradation for other UEs. Thus, Est is computed as a sum over k steps, representing estimated duration of the data transmission. The actions in the MDP for selection of the communication path are defined as: 1) transit to another state s' (change current path) if it improves M_q or 2) stay in the current state s (use the same path) if M_q cannot be improved by selection of another path. However, with dynamicity of the mobile networks, transitions from one state to another (option 1) cannot be stationary mapped to states and need to reflect changes of the network topology and transmission parameters of each link. Thus, we calculate table of transitions among states every time when the task is offloaded. Optimum policy π is obtained at the end of the algorithm and it gives desired policy maximizing the reward. As the delay and the energy are used as metrics, the optimal policies can be calculated in order to minimize delay, energy or a trade-off between both metrics. The reward depends on the delay due to handover (t_{H}) if the handover is performed, delay by the transmission over radio (t_{R}) and transmission delay on backhaul (t_{B}).

Thus, the reward for transition from the state s to the s' defined in (5.1) can be rewritten for purposes of the path selection as follows.

The transmission delays t_{R} and t_{B} are computed knowing amount of data to be transferred over radio ($v_{bits}^{R,i}$) and backhaul ($v_{bits}^{B,i}$) and knowing capacity of the radio link (c_i^R), capacity of backhaul of the serving (c_i^B) and the computing (c_x^B) cells:

$$t_{\mathrm{R}} = \frac{v_{bits}^{R,i}}{c_i^R} \tag{5.6}$$

$$t_{\mathrm{B}} = \frac{v_{bits}^{B,i}}{c_i^B} + \frac{v_{bits}^{B,i}}{c_x^B} \tag{5.7}$$

5.2.4 Path selection algorithm

The pseudo-code for the proposed algorithm, which selects the path between the UE and the computing SCeNBs for given γ is shown in Algorithm 4. The algorithm calculates delay t_q and energy E_q spent by the UE for delivery of the offloaded task to each $SCeNB_x$ (Step 3) using available radio links of neighboring cells (Step 4). Delay and energy due to

transmission using radio of the $SCeNB_i$ to deliver data to the $SCeNB_x$ are derived using (5.6), (5.7) and (A.4), (A.11), respectively (Steps 6 and 7). If data is sent over backhaul link (Step 8), its delay is added to the path delay (Steps 9, 10). Afterwards, the delay and energy of each combination of radio and backhaul links is calculated (Steps 15 and 16). Impact of handover on the path selection is included by adding delay of handover (t_{H}) and energy consumed by the UE during handover (E_H) to the delay and energy derived for the $q-th$ path (Steps 18 and 19). Energy consumed by the UE during handover is calculated using (A.4) by substituting t_{H} for t_{R}. Subsequently, t_q and E_q are normalized in order to be weighted (Steps 23 and 24). Then, the path metric M_q is calculated by weighting $\overline{t_q}$ and $\overline{E_q}$ (Step 25). Finally, the new path $q'_{s'}$ with the lowest M_q for given γ is returned (Steps 27, 28).

Algorithm 4 Selection of path for data delivery

c_i^B,c_x^B, c_i^R,$v_{bits}^{R,i}$,$v_{bits}^{B,i}$,γ

1: $t_q \leftarrow null$
2: $E_q \leftarrow null$
3: **for** $x \in X$ **do**
4: **for** $i \in I$ **do**
5: $t_{\mathrm{R}} \leftarrow v_{bits}^{R,i}/c_i^R$
6: $d_x^i \leftarrow t_{\mathrm{R}}$
7: $e_x^i \leftarrow E[t_{\mathrm{R}}]$
8: **if** $v_{bits}^{B,i} > 0$ **then**
9: $t_{\mathrm{B}} \leftarrow v_{bits}^{B,i}/c_i^B + v_{bits}^{B,i}/c_x^B$
10: $d_x^i \leftarrow d_x^i + t_{\mathrm{B}}$
11: **end if**
12: **end for**
13: **end for**
14: **for** $q \in Q$ **do**
15: $t_q \leftarrow \sum_q d_x^q$
16: $E_q \leftarrow \sum_q e_x^q$
17: **if** handover **then**
18: $t_q \leftarrow t_q + t_{\mathrm{H}}$
19: $E_q \leftarrow E_q + E_H$
20: **end if**
21: **end for**
22: **for** $q \in Q$ **do**
23: $\overline{t_q} \leftarrow t_q/max\{t_q\}$
24: $\overline{E_q} \leftarrow E_q/max\{E_q\}$
25: $M_q \leftarrow \gamma\overline{E_q} + (1-\gamma)\overline{t_q}$
26: **end for**
27: $q'_{s'} \leftarrow argmin\{M_q\}$
28: **return** $q'_{s'}$

5.2.5 Implementation aspects

To enable implementation of the proposed path selection algorithm, capacity of radio link, capacity of backhaul link, and transmission/reception power level at the UE's side must be obtained. The capacity of uplink radio link is derived from the number of allocated RBs. This can be obtained through uplink grant reception [42]. Similarly, the capacity of downlink radio link depends on the number of allocated RBs. This information is derived by means of downlink assignment as described in [42]. Apart from the number of allocated RBs, the capacity depends on MCS used for transmission based on SINR. From knowledge of the amount of bits to be transmitted to each computing $SCeNB_x$, and radio capacity of each SCeNB, we calculate the delay of radio transmission using (5.6). The capacity of backhaul connection is calculated based on known maximum backhaul capacity and link utilization. Required parameters for calculation of energy consumption are measured directly by the UE or obtained via control channels from the SCeNBs.

To determine the most suitable paths, it is necessary to identify cells, which are in communication range of the UE. This can be done according to the SINR. In the common mobile networks, such as LTE-A, the UE can monitor SINR from the SCeNBs included in Neighbor Cell List (NCL) (for more details about NCL, refer to [166]). The NCL contains all potential neighbors of the UE's serving cell. Thus, this corresponds to the list of the SCeNBs, which might be available for data transmission.

Each SCeNB can be switched off at any time since the SCeNBs can be deployed also by users (e.g., femtocells) [167]. Thus, a secondary path should be defined for a case when the primary path is no longer available due to its failure. To keep routing overhead low in case of the link failure, data to be sent over this link will be rerouted through the original serving SCeNB selected according to signal quality, if possible. In case of the serving SCeNB failure, the UE will reinitiate path selection as there is a major change in state of links and there is no other backup route. Note that this problem requires also selection of a new serving cell for communication purposes. However, this is a common problem, for which existing mobile networks are able to find a solution by selection of new serving cell according to RSSI (for more details, see [168]).

5.2.6 Complexity of the path selection algorithm

Complexity of the proposed path selection algorithm is proportional to the number of computing SCeNBs (n) and the number of SCeNBs in radio communication range of the UE (m). The number of possible paths can be computed as partial permutation. Thus, the complexity of algorithm is $O(m^n)$. This complexity might be redundant as many SCeNBs in communication range of the UE provide radio channel of quality not suitable to satisfy requirements on delay imposed by a service (t_{req}). Hence, we narrow-down former set of all SCeNBs in communication range of the UE (I) to the set Y with a size of m_y, consisting of the SCeNBs with SINR above a threshold ρ_{SINR}. The set Y is created to cut off unusable SCeNBs with very low channel quality. The cut, defining the set Y, has no negative impact on performance of the proposed algorithm, as the cut

removes only the SCeNBs, which cannot be used for communication due to low channel quality. This means that we set the SINR threshold ρ_{SINR} equal to a minimum SINR when devices can communicate and thus, SCeNBs with $SINR_{SCeNB} < \rho_{SINR}$ are not considered for the path selection. note that the size of set Y can be controlled by setting the threshold ρ_{SINR}. However, high ρ_{SINR} could lead to performance degradation as some base stations in proximity of UE, which potentially available for data transmission, might be excluded from the set Y disregarding the amount of available radio resources and backhaul. Anyway, considering low radius of small cells, wall attenuation and interference, the number of small cells in radio communication range (i.e., $SINR_{SCeNB} > \rho_{SINR}$) is very low (in our simulations, typically between two and four). Thus, complexity of the proposed solution after removing unsuitable base stations form set Y is also kept at low level.

Consequently, the set Y includes only SCeNBs, which can provide SINR high enough to satisfy t_{req}, i.e., the set Y is defined as:

$$Y = \{y \mid y \in I, y > \rho_{SINR}\} \tag{5.8}$$

By this approach, the list of SCeNBs in UE's proximity is reduced from a size of m to m_y. Consequently, the complexity of the proposed path selection algorithm is reduced from $O(m^n)$ to $O({m_y}^n)$, where $m > m_y$. Note that this leads to replacement of the set I by the set Y in Algorithm 4 in Step 4.

5.2.7 Evaluation Methodology and Scenario

In this section, scenarios, deployment and simulation models used in MATLAB simulations for performance evaluations are presented. The simulation area is composed of two-stripes of buildings (as shown in Figure 5.4) as suggested by 3GPP in [157]. The size of each building's block is 20 x 100 m and blocks are separated by streets with a width of 10 m. The overall simulation area is composed of 4 x 4 blocks of offices or apartments. The size of the whole simulated area is 430 x 270 m. Fifty outdoor UEs are randomly deployed at the beginning of the simulation and they move along the streets according to Manhattan Mobility model [169], with a movement speed of 1 m/s. In addition, also indoor UEs are randomly deployed in offices with 20% offices occupied with one UE, i.e., there are 64 indoor UEs. Movement of the indoor UEs is modeled so that the UEs move within the apartments at discrete positions with a specific time distributions as defined in [170]. Inside the buildings, also the SCeNBs are randomly dropped to the offices with equal probability in a way that 20% of offices are equipped with a SCeNB. Therefore, 64 SCeNBs are deployed indoor. Besides the SCeNBs, also a macrocell eNB is placed outside the block of buildings at coordinates of [425 m, 265 m] (see Figure 5.4).

The offloaded tasks is computed at 1, 2, 3 or 4 SCeNBs, with equal probability of each option. One of the computing SCeNBs is the serving one if this one can offer enough computing resources as suggested in [19]. In simulations, the computing SCeNBs are selected as a random set of n closest available SCeNBs.

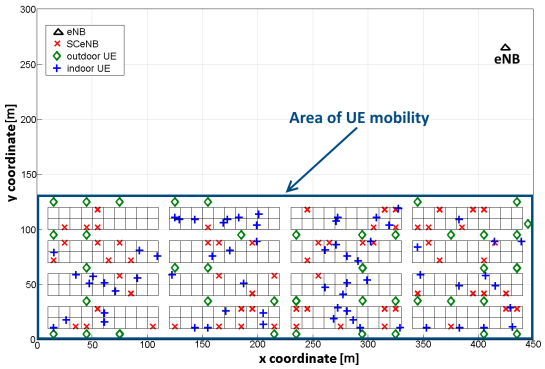


Figure 5.4. Simulation scenario with example of deployment of buildings, users, SCeNBs and eNB for simulations.

We assume the size of offloaded task is either 300 kB or 30 MB to represent two different loads corresponding to different types of applications [171]. Interval between two offloaded tasks is set to 64 s and 512 s for a size of tasks of 300 kB and 30 MB, respectively. This interval is selected to generate enough traffic so that one request transfer affects selection of path for another.

If two or more offloaded tasks are generated at the same moment, the path selection is done sequentially for each task. Therefore, the impact of the path selection for the first offloaded task is subsequently considered for the second offloaded task and so on.

Major parameters of the simulation, summarized in Table 5.1, are in line with the recommendations for networks with small cells as defined by 3GPP in [157]. We also follow parameters of the physical layer frame structure for LTE-A mobile networks and signal propagation as defined in the same document. Based on [109] and [172], we set handover delay to be 30 ms. In LTE-A, OFDMA is used for communication at the physical layer in downlink whereas SC-FDMA is used in uplink. The smallest unit to be allocated to the UE is a RB, which consists of 12 subcarriers and 7 symbols. Downlink and uplink are separated by means of Frequency Division Duplex (FDD).

Radio and backhaul resources are shared among the UEs in such manner that newly incoming request can be assigned with up to half of available resources to guarantee resource availability also for other potential UEs and services. A part of radio link capacity is assumed to be consumed by the background traffic (common voice and data services exploited by other users). Thus, the maximum number of available RBs per subframe for uplink and downlink in our simulations is 40 and 80, respectively.

We consider SCeNBs connected to the operator's network through either DSL or optical fiber. Maximum throughput of both is generated by a normal distribution with mean value μ and standard deviation σ as specified in Table 5.2. The optical fiber is used solely for the corporate scenario which assumes several cells within one building [170] sharing the same backhaul. All cells belonging to the same corporate building are interconnected with Local Area Network (LAN) offering throughput of 100 Mbit/s among the SCeNBs. If two or more SCeNBs within the same corporate building communicate with each other, we assume direct communication between the SCeNBs via LAN. Consequently, no part

Table 5.1. Simulation parameters

Parameter	**Value**
Simulation area	430 x 270 m
Carrier frequency	2 000 MHz
Bandwidth for downlink/uplink	20/10 MHz
Tx power of eNB/SCeNB	43/23 dB
Attenuation of external/internal/separating walls	20/3/7 dB
SCeNB deployment ratio	0.2
Shadowing factor	6 dB
Handover interruption duration	30 ms
Number of Indoor UEs/Outdoor UEs/SCeNBs	64/50/64
Speed of outdoor users	1 m/s
Traffic generated by one request	300 kB/30 MB
Time between two requests for 300 kB/30 MB tasks	64/512 s
Simulation time	20 000 s
Number of simulation drops	4

of the offloaded task is distributed to the core network over optical fiber. The DSL backhaul connection corresponds to the residential scenario where the SCeNBs are deployed in private flats and connected to the core network [170]. For both scenarios, the UE can communicate with the computing SCeNBs also via eNB. The eNB is connected to operator's network through a link with a throughput of 1000 Mbit/s.

Table 5.2. Parameters of backhaul models.

Parameter	**Value**
Optical fiber μ (uplink/downlink)	100/100 Mbit/s
Optical fiber σ (uplink/downlink)	11.5/11.5 Mbit/s
DSL μ (uplink/downlink)	1/5.5 Mbit/s
DSL σ (uplink/downlink)	100 Mbit/s
eNB uplink/downlink	1000/1000 Mbit/s

5.2.8 Simulation Results

In this section, simulation results are presented and discussed. The performance is evaluated for the proposed PSwH algorithm and also for commonly adopted SO approach [19,21,113,173] (only the serving SCeNB selection based on RSSI level is assumed). Simulation results are divided into subsections analyzing : 1) transmission delay, 2) energy consumed by the UE, 3) satisfaction of users with experienced delay, 4) load of the SCeNB's backhaul, and 5) number of additional handovers generated by the PSwH. We also discuss selection of proper values of weighting parameter γ and we summarize major findings from simulations in this section.

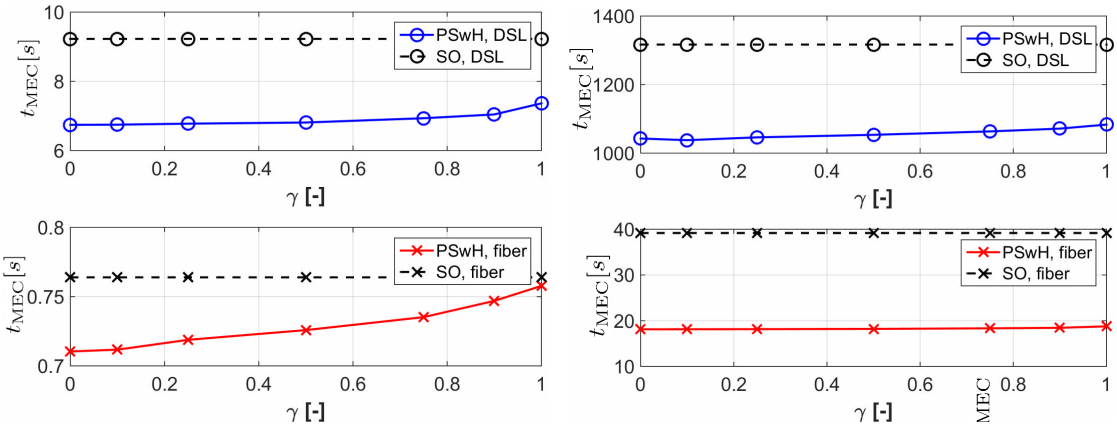


Figure 5.5. Average time (t_{MEC}) required for transmission of offloaded task with size of (a) 300 kB and (b) 30 MB for DSL backhaul (top subplot) and fiber optic (bottom subplot) backhauls.

Delay of UE data transmission/reception

Impact of the PSwH algorithm on the average delay caused by transmission of the offloaded task between the UE and the computing SCeNBs is depicted in Figure 5.5a and Figure 5.5b. From both figures, we can observe that delay increases with γ. This is because high γ indicates priority for low energy consumption while delay becomes less important (see (5.3)). The proposed PSwH reaches lower delay comparing to the SO for all values of γ. Low delay achieved by the PSwH results from avoiding low quality backhaul if it is possible to transmit data directly to the computing SCeNBs or through different SCeNBs with less loaded backhaul. For the offloaded task with a size of 300 kB, the average transmission delay t_{MEC} is reduced by up to 26.9% for DSL backhaul and up to 7% for optical fiber backhaul, as shown in Figure 5.5a. For the offloaded task's size of 30 MB, the PSwH shortens the delay by up to 21.5% comparing to the SO in case of DSL backhaul and by up to 53.7% for the optical fiber backhaul as shown in Figure 5.5b.

Energy consumed by UE for data transmission/reception

The proposed PSwH should avoid draining of the UE's battery caused by data transmission/reception and handover. By increasing γ, radio paths with lower energy consumption are used more often and the energy consumption is decreasing (see Figure 5.6a and Figure 5.6b). Energy required for the UE's transmission depends on transmission power level and transmission duration as specified in (A.4). Lower energy consumption comparing to the SO is achieved by shortening the transmission time resulting from performing handover to less loaded SCeNBs (offering more available RBs for transmission). The reason for lower energy consumption for transmission via less loaded SCeNB is that a linear increase in the number of consumed RBs leads to a linear decrease in the transmission delay while increase in the energy consumption is logarithmic (see (A.3)). Another reason for lowering the energy consumption by the PSwH is the usage of the connection

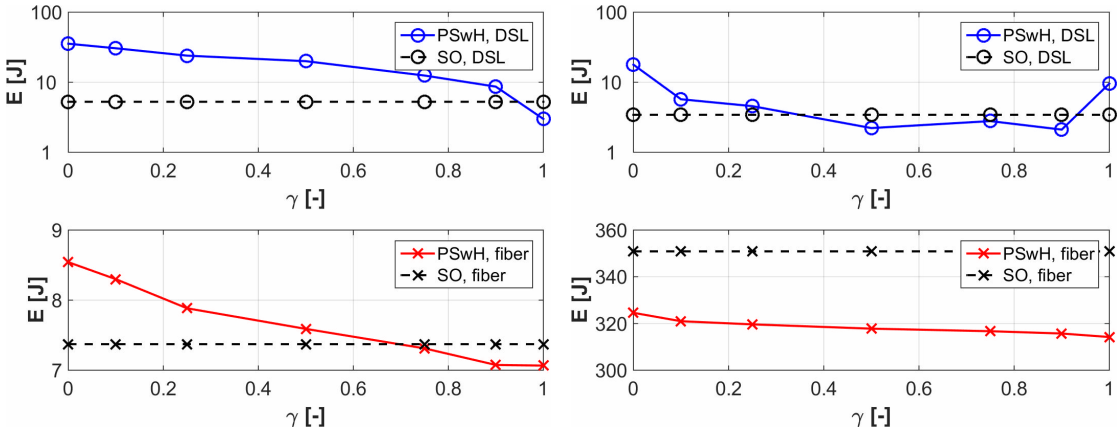


Figure 5.6. Average energy (E) required for transmission of offloaded task with size of (a) 300 kB and (b) 30 MB for DSL backhaul (top subplot) and fiber optic (bottom subplot) backhauls.

with a more robust MCS, which requires lower transmission power (see [162] for more details).

From Figure 5.6a and Figure 5.6b, we can see that the PSwH reduces energy consumption by up to 3.2% in case of the DSL backhaul and by up to 4.1% for the optical fiber backhaul if the offloaded task is of 300 kB as shown in Figure 5.6a. For the offloaded task of 30 MB (shown in Figure 5.6b), the PSwH lowers energy consumption by up to 4.1% for DSL backhaul and by up to 10.4% for the optical fiber. The energy consumption can be increased comparing to the SO if the users do not care about energy (low γ). However, in this case, the users indicate their preference for the delay so they are not unhappy with increased energy consumption.

There is one singular point when impact of γ on the energy is unexpected (energy rises with γ). This situation is shown in Figure 5.6b for large offloaded tasks (30 MB) and low quality backhaul (DSL). In this scenario, the algorithm is trying to minimize energy consumption by selection of the most appropriate radio path disregarding delay (see (5.3)). Hence, the algorithm tends to associate all UEs to the SCeNBs with radio links requiring the lowest energy consumption. However, for the UEs trying to offload data later when other transmissions are already in progress, not enough radio resources are available. Consequently, those UEs are associated to the SCeNBs, which may lead to even higher energy consumption than in case of the SO.

Satisfaction of users with experienced transmission delay

The satisfaction of UEs with experienced transmission delay t_{MEC} with respect to their required delay t_{req} is shown in Figure 5.7 and Figure 5.8 for DSL and optical fiber backhauls, respectively. The satisfaction is understood as a ratio of users (R_s), who experience delay lower than the requested one (i.e., $t_{\mathrm{MEC}} \leq t_{\mathrm{req}}$). The satisfaction increases as γ decreases since lowering delay becomes of higher priority than energy consumption.

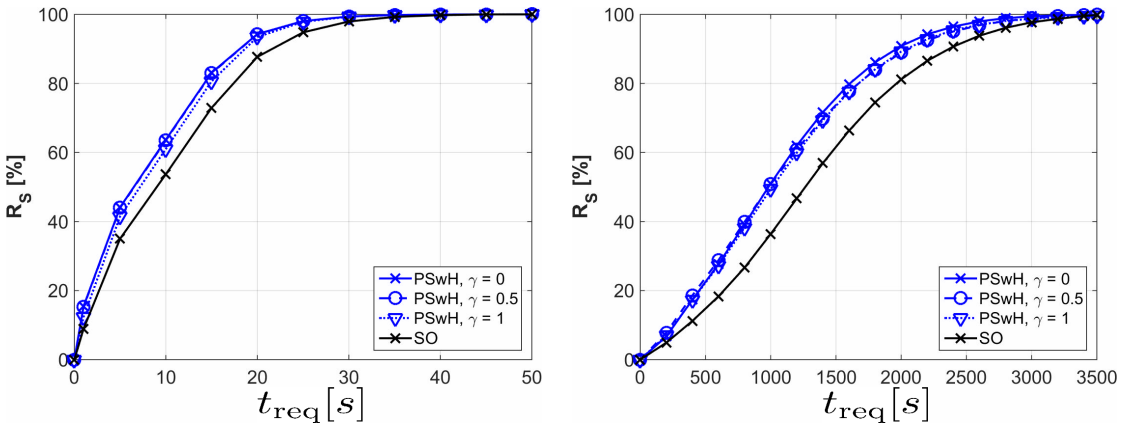


Figure 5.7. Ratio of users satisfied with experienced delay, R_S, for DSL backhaul for offloaded task of 300 kB (a) and 30 MB (b).

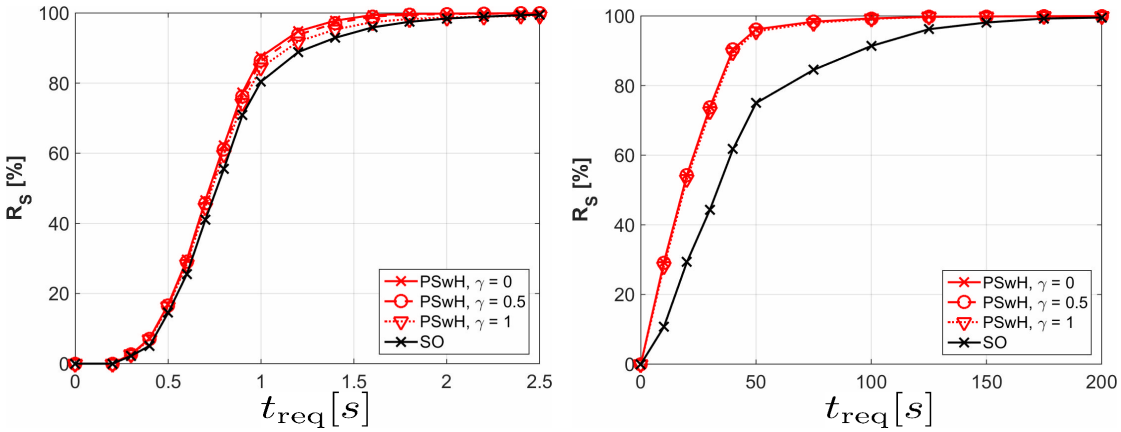


Figure 5.8. Ratio of users satisfied with experienced delay, R_S, for optical fiber backhaul, for offloaded task of 300 kB (a) and 30 MB (b).

As can be seen from Figure 5.7 and Figure 5.8, the UEs' satisfaction is increasing with t_{req} for both compared algorithms. This fact is expected as more time is available for delivery of data for higher t_{req}. Comparing the PSwH with the SO for DSL backhaul, the proposed algorithm increases the satisfaction up to 10% for the offloaded task with a size of 300 kB (Figure 5.7a) and up to 15% for the offloaded task with a size of 30 MB (Figure 5.7b).

For the optical fiber backhaul, the satisfaction of UEs with experienced transmission delay is shown in Figure 5.8. The PSwH improves the satisfaction by up to 7% for the offloaded task of 300 kB (Figure 5.8a), and up to 29% for the offloaded task of 30 MB (Figure 5.8b).

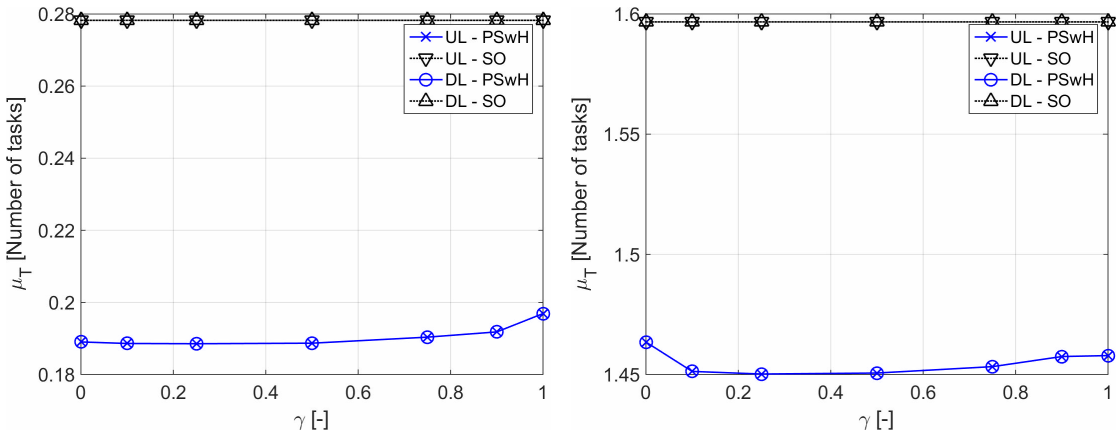


Figure 5.9. Mean number of the offloaded tasks, μ_T, transmitted over DSL backhaul for the offloaded task size of 300 kB (a) and 30 MB (b).

Load of small cell's backhaul

The proposed algorithm takes advantage of handovers to speed up data delivery and also to offload the backhaul of SCeNBs. The load of backhaul (μ_T) is represented by a mean number of the offloaded tasks transmitted per backhaul link and time. For the PSwH, the backhaul load increases with γ since a priority is given to lowering the UE's energy consumption while backhaul capacity (represented by transmission delay) is of a lower priority. This behavior results from the lowering energy consumption (high γ), which leads to selection of less energy consuming radio link even if backhaul has to be used. Contrary, the most of the traffic is transmitted directly to the computing SCeNB in radio communication range if the users prefer low delay (low γ).

In Figure 5.9a, we can see that the PSwH reduces the DSL backhaul load by up to 32% comparing to the SO for both uplink and downlink for the offloaded task of 300 kB. For the offloaded task of 30 MB, more than 9% decrease in DSL backhaul utilization is observed as well for both directions as shown in Figure 5.9b. The decrease in backhaul load by the PSwH is due to exploitation of the radio link rather than low quality backhauls.

In case of the optical fiber, the PSwH lowers the backhaul load by up to 11% for the offloaded task with a size of 300 kB and by up to 15.5% for the offloaded task with a size of 30 MB, as shown in Figure 5.10a and Figure 5.10b, respectively.

Number of performed handovers caused by the proposed algorithm

The ratio of additional handovers introduced by the PSwH for the transmission of offloaded task (R_H) is shown in Figure 5.11. If the PSwH is used, the number of handovers for delivery of the offloaded task of 300 kB is increased by 53-56% for the optical fiber backhaul and by 53-55% for the DSL backhaul (see Figure 5.11a). If the size of offloaded task is 30 MB (Figure 5.11b), the number of handovers is increased by 5% for DSL backhaul and by 12.5% for optical fiber backhaul. For both backhauls, the number of

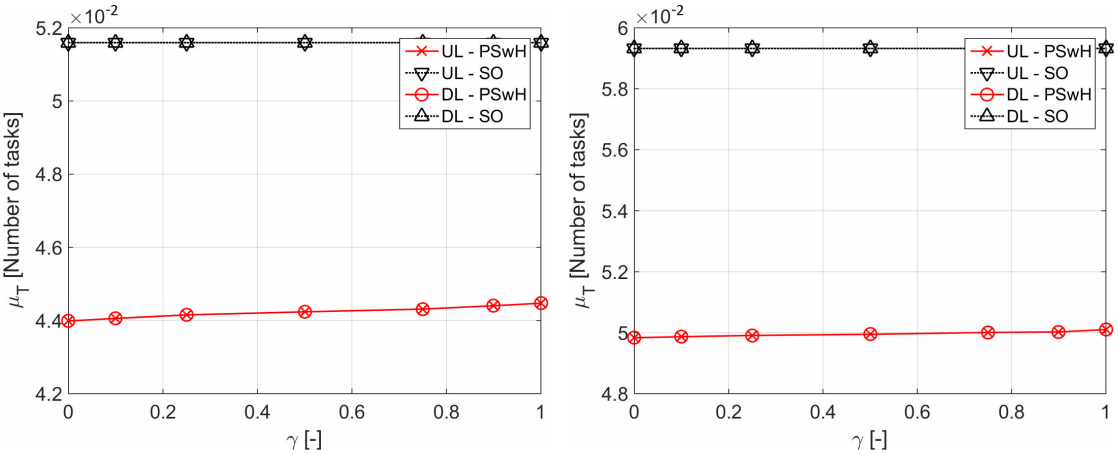


Figure 5.10. Mean number of the offloaded tasks, μ_T, transmitted over optical fiber backhaul for the offloaded task size of 300 kB (a) and 30 MB (b).

handovers increases with γ for low values of γ and then decreases for high values of γ. This behavior is a result of combination of handovers initiated for minimization of the energy consumption as well as for minimization of the delay for $0 < \gamma < 1$. For $\gamma = 0$ or $\gamma = 1$, the handover is initiated less often as only either energy consumption or delay are targeted. Note that the impact of γ on the number of additional handovers is very low (below 3.5%).

Less significant increase in the number of handovers for large offloaded tasks (30 MB) is caused by more time required for transmission of such task. Therefore, the radio links of the SCeNBs in communication range of the UE (included in set I) are heavily loaded for a longer period of time. Consequently, allocation of resources at the overloaded neighboring SCeNBs for the users associated to another cell is not feasible.

The proposed algorithm introduces additional handovers, which can lead to redundant signaling and interruption in communication due to performing handover (known as handover interruption). The signaling overhead generated per handover is in order of kb [174]. The overall number of handovers per one offloaded task is very low (roughly 0.8 in average). Hence, total handover overhead is in order of kb per offloaded task and can be considered negligible. The second problem, handover interruption, is not related to the SCC services as the users do not care about interruption in transmission of the offloaded task if the computation results are delivered within desired delay t_{req}. The handover interruption is considered in the PSwH algorithm (see (5.1)). Thus, all above-presented results already include impact of the handover interruption. Of course, the handover interruption introduced by the PSwH can degrade quality of conventional services (voice, video, etc.) running at the UE simultaneously with offloaded SCC services. In case of simultaneous usage of the SCC offloading service and common real-time service, the user must indicate priority for one type of services. If the preference is given to the SCC service, the user should not be disappointed with lower quality of the secondary service. Contrary, if the preference is given to the conventional real-time (non-SCC) service, the SCC service will

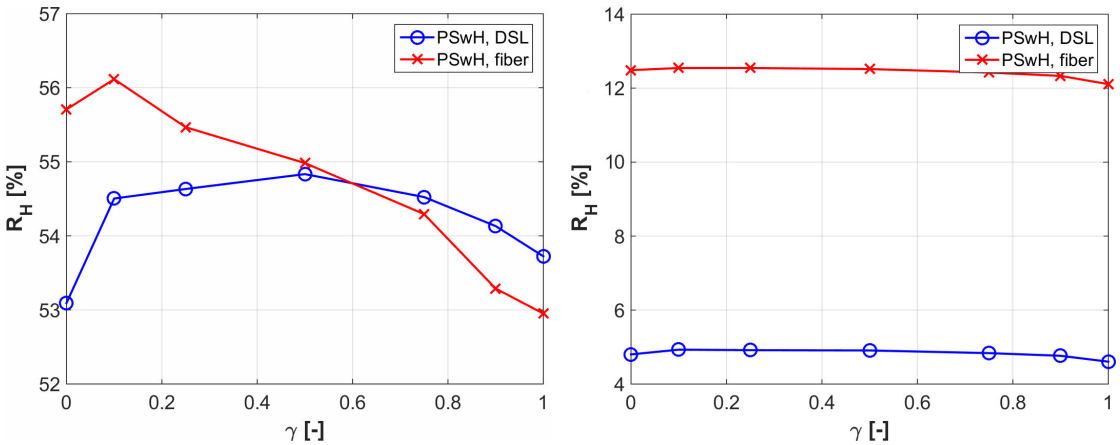


Figure 5.11. Ratio of additional handovers generated by the PSwH algorithm, R_H, with respect to the SO for offloaded tasks of 300 kB (a) and 30 MB (b).

be handled in conventional way (i.e., by means of SO algorithm) with no gain in delay or energy consumption but also with no degradation in QoS for the non-SCC service. Note that the SCC is intended mainly for delay sensitive and real-time services (applications). Therefore, simultaneous usage of the SCC service and common non-SCC service is not very likely.

Discussion of results and selection of proper γ

In this section, proper selection of γ for the proposed algorithm is discussed along with gain in above-mentioned performance metrics introduced by the PSwH.

The proper γ is selected in such a way that the delay reduction comparing to the SO is maximal while energy consumption is still lowered or at least not impaired comparing to the SO. The selected values of γ are shown in Table 5.3. The proper value of γ spans over the whole range (i.e., from 0 to 1) and individual proper value depends on combination of backhaul quality and a size of the offloaded task. Consequently, also the gain (Δ) introduced by the PSwH comparing to the SO varies for backhaul types and a size of the offloaded tasks. The gain Δ is defined as improvement introduced by the PSwH with respect to the SO for each performance metric. For example, the gain in delay is defined as $\Delta t_{\mathrm{MEC}} = (t_{PSwH} - t_{SO})/t_{SO}$. Therefore, the negative numbers (green color) in this table represent improvement introduced by the PSwH comparing to the SO (e.g., the PSwH reduces delay by 20.2%) while the positive numbers (red color) indicate worsened performance (additional handovers introduced by the PSwH).

The variation of gain for different backhauls and offloaded task size is caused by availability of each backhaul for data transmission and its ability to handle given level of load introduced by the offloading tasks. For high quality optical fiber backhaul, the small tasks (300 kB) can be handled even by the SO algorithm as the optic is able to distribute such small amount of data easily. Thus, to reach a gain by the PSwH, a high number of handovers must be performed to find more suitable way of data distribution. However,

for other scenarios optical fiber with large tasks or DSL with both sizes of the tasks), the SO fails in distribution of the tasks over backhaul, which can be easily overloaded by the offloaded tasks. Consequently, the gains introduced by the PSwH become more significant.

Table 5.3. Summarized improvement (green color) in performance metrics introduced by the PSwH comparing to the SO for proper values of γ.

Backhaul type	**Size of task**	**Proper γ**	Δt_{MEC} [%]	ΔE [%]	$\Delta \mu_T$ [%]	R_H [%]
DSL	300 kB	1	-20.2	-3.2	-29.2	+53.5
DSL	30 MB	0.5	-19.1	-0.9	-9.1	+5.1
optical fiber	300 kB	0.75	-3.8	-0.8	-14.3	+54.3
optical fiber	30 MB	0	-54.3	-7.5	-16	+12.4

5.2.9 Conclusion

In this section, we have proposed a new path selection algorithm for delivery of the offloaded tasks between the UE and the cloud-enhanced small cells, representing MEC hosts. The algorithm forces the UE to perform handover if it is efficient in terms of the overall transmission delay (considering radio and backhaul) and/or energy consumption of the UE. In order to find a trade-off between transmission delay and energy efficiency, weighting of both metrics is introduced. The proposed algorithm reduces the transmission delay by up to 20.2% and 54.3% in scenario with small cells connected to the operator's network by the DSL backhaul and optical fiber, respectively. At the same time, the energy consumption of the UE can be lowered by 3.2% and by 7.5% for DSL and optical fiber backhauls, respectively. Notice that the improvement accomplished by the PSwH depends on the size of offloaded task together with used backhaul connection. The proposed algorithm also increases user's satisfaction with experienced delay (up to 29%) and lowers backhaul load (up to 32%). The improvements reached by the proposed algorithm are at the cost of additional handovers. Nevertheless, delay introduced by these additional handovers is already considered in the path selection algorithm. Therefore, the handovers do not decrease QoS but leads only to negligible additional overhead (few kb per offloaded task).

As the algorithm can select efficient path for downlink and uplink independently, it is suitable also for mobility management of moving user's exploiting the SCC services.

5.3 Joint computation and communication resource allocation under fixed prediction accuracy

In the previous section, we have proposed an algorithm for communication resource allocation (communication path selection). However, due to the inherited mobility of

the UEs, it is necessary to determine allocation of the computation resources as well. Therefore, in this section, we describe a solution for joint communication and computation resource allocation.

This section is organized as follows, in the next section, we define model of the investigated MEC system. In Section 5.3.2, the proposed algorithm is described. Simulation environment and results are presented in Section 5.3.3, and Section 5.3.4 concludes the joint communication and computation resource allocation.

5.3.1 System model

The system is assumed to be composed of set S of base stations $s \in S$. To generalize the system model, the base station can be represented either by macro cell (eNB), SCeNB, or Femto Cell eNB (HeNB). Unless otherwise stated, the label eNB covers all types of base stations. For each UE, the serving eNB $s' \in S$ is selected as the one with the highest Received Signal Strength (RSS). As the UE moves, the serving eNB is updated by following a conventional hard handover procedure (see [163]). The handover introduces an interruption in communication with a duration of t_{HO} (known as handover delay or handover interruption). This delay consists of time required to break the connection with current serving eNB and to establish a connection with a new eNB. Note that the UE can neither transmit nor receive data during hard handover in mobile networks.

To facilitate MEC, a VM for the UE is created at a base station, which is denoted as $s_{VM} \in S$. Number of the UEs with VM allocated at the s-th eNB is labeled as $n_s^{VM}(t)$, whereas number of UEs utilizing communication resources of the s-th eNB is denoted as $n_s^c(t)$. The VM can be placed at i) the eNB selected based on offloading delay or energy and kept there [21], ii) the serving eNB and migrated to a new serving eNB after handover [175], iii) dynamically placed by considering the UE's movement [130].

The offloaded task is defined by its size L (in bits), the number of instructions to be processed B, and the size of computed results R (in bits). As a possibility to migrate the VM is considered, the size of the migrated VM is defined as a number of bits G.

In Figure 5.12, an example of UE's movement for two adjacent time instances t and $t + \Delta t$ is depicted. The UE communicates with the serving eNB via radio channel with $SINR_s$ and capacity c_s^R. Each eNB is connected to mobile operator's core network (network connection through which the eNB is connected to the Internet) via backhaul with capacity c_s^B.

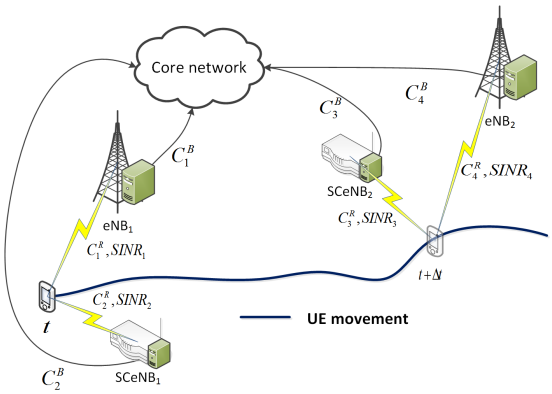


Figure 5.12. System model.

A set of eNBs with which the UE can communicate over radio channel is denoted as I; $I \subset S$. The set I includes the eNBs for which the $SINR$ observed by the UE is above $SINR_{min}$. An example of $SINR_{min}$ for LTE-A network and Block Error Rate (BLER) of 10% is a value of -6.9 dB [176]. If the UE needs to deliver an offloaded task to s_{VM}, the transmission can be done directly via radio if $s' = s_{VM}$ or the offloaded task is transmitted via radio of s' and then via backhaul connection between s' and s_{VM}. The capacity available for delivery of the offloaded task from the UE to the eNB with allocated VM is calculated as:

$$c_{UE,s_{VM}} = \begin{cases} c^R_{s'} & s' = s_{VM} \\ min\{c^R_{s'}, c^B_{s',s_{VM}}\} & otherwise \end{cases} \tag{5.9}$$

where capacity between two eNBs s' and s_{VM} is calculated as $c^B_{s',s_{VM}} = min\{c^B_{s'}, c^B_{s_{VM}}\}$.

As prediction is considered in our model, predicted communication capacity and available computing capacity (in instructions per second) are denoted as $\tilde{c}(t)$ and $\tilde{k}(t)$, respectively, where t is the time instance at which capacity and computational resources are being predicted. The predicted radio capacity is derived from predicted position of the UE mapped to SINR maps as introduced in [177]. From information about SINR and $n^c_s(t)$, available capacity of radio is computed as:

$$\tilde{c}^R_s(t) = thr\{MCS\{SINR_s(t)\}, \frac{n^{RB}_s}{n^c_s(t)}\} \tag{5.10}$$

where $MCS\{SINR_s(t)\}$ maps SINR to MCS (e.g. by [176]), n^{RB}_s specifies the number of all RBs of the s-th eNB and function $thr()$, maps MCS and the number of RBs to the number of bits for transmission as described in [31]. For backhaul, the predicted capacity is calculated as $\tilde{c}^B_s(t) = \frac{\bar{c}^B_s}{n^c_s(t)}$, where $\bar{c}^B_s$ denotes backhaul capacity of the s-th eNB. Apart from the capacities, the predicted delay of offloading consists of delay due to uploading the offloaded task to the VM, $t^s_{\mathrm{O}} = \frac{L}{\tilde{c}_{UE,s_{VM}}(t)} + \sum t_{\mathrm{HO}}$, computation delay, i.e., time required to process the offloaded task by the VM, $t^s_{\mathrm{P}} = \frac{B}{\tilde{k}(t)}$, delay due to collecting results by the UE from the VM, i.e., downloading computed results from the eNB where the VM is allocated, $t^s_{\mathrm{C}} = \frac{R}{\tilde{c}_{s_{VM},UE}(t)} + \sum t_{\mathrm{HO}}$, delay of the VM migration represented by time required to copy and start the VM from current serving eNB to a new serving eNB,

$t^s_{\mathrm{M}} = \frac{G}{\bar{c}_{s,s_{VM}}(t)}$ and delay of starting VM instead of migrating VM, t_{APP}.

As we target offloading of real-time applications, we assume that the VM is pre-allocated [178]. Thus, delay due to starting the VM is equal only to delay of starting the offloaded application on the side of the VM. Total delay of one offloaded task is then defined as $t^s_{\mathrm{MEC}} = t^s_{\mathrm{O}} + t^s_{\mathrm{P}} + t^s_{\mathrm{C}} + \sum t^s_{\mathrm{M}} + \sum t_{\mathrm{APP}}$ and it is a sum of communication, computation, VM migrations, and starting of VMs.

5.3.2 Dynamic Resource Allocation

The proposed dynamic resource allocation in this section is based on our previous PSwH algorithm described in [162], which exploits reward function fromMDP to select the communication path q. The PSwH forces the UE to perform handover to new eNB if it is profitable for the UE from the offloading point of view. In this section, we enhance the PSwH algorithm by mobility prediction and we design a cooperative algorithm for dynamic VM placement based on calculating reward in terms of communication capacity and incorporating load balancing. Both algorithms are based on reward function from MDP and utilize prediction window denoted by τ.

The idea of cooperation between both proposed algorithms is to dynamically place the VM before the UE starts offloading, as migration (or start) of the VM if offloading is in progress would increase the offloading delay by the VM migration (ot starting the VM). Therefore, when the UE starts offloading its task, the VM will be already prepared at the suitable eNB. From the perspective of delay, the suitable eNB would be the eNB with good radio channel (high SINR), as the channel quality directly relates to the communication channel capacity. Also the UE exploits the PSwH algorithm enhanced with mobility prediction to select a suitable communication path (i.e., the serving eNB) in order to further reduce offloading delay. Cooperation is achieved by starting the algorithm for dynamic VM placement in-between offloading of two consecutive tasks, when certain radio conditions are met (SINR is below a given threshold). Both, the PSwH enhanced by mobility prediction and the dynamic VM placement algorithms are based on MDP, as described in Section 5.1. In the MDP we replace states s and s' with selected communication path q and q' respectively. With respect to PSwH in [162], the reward function is based on prediction, i.e., estimation Est is replaced by prediction $Pred$.

As the VM have to be ready to process the offloaded task when offloading starts [98], the decision on VM placement should be made before the offloading starts. Therefore, the algorithm for dynamic VM placement is initiated if eNBs with $(SINR_s > SINR_{s_{VM}} | s \in S, s \neq s_{VM})$ are in communication proximity of the UE in order to provide sufficient capacity of radio communication channels.

To find the best placement of VM, $SINR$ to set S is predicted. However, the set S could be quite large. Thus, in our proposal, we define a reduced set $Z\{z \in Z | (SINR_z > SINR_{min}) \cap (n^{VM}_z(t) < n_{limit})\}$. In this set, each eNB z has $SINR$ above $SINR_{min}$. Also, the set includes only the eNBs, which are not overloaded, i.e., their load is below n_{limit}, to distribute computational load more equally.

The proposed algorithm for dynamic VM placement is described in Algorithm 5. For each eNB in Z (step 1) and each eNB from set I (step 2), $SINR$ is predicted by applying SINR map [177] on predicted UE's mobility (step 3). Communication capacity is predicted from predicted $SINR$ and $n_s^c(t)$(step 4). In order to prefer eNBs with good channel quality in the future (next time steps), a slope of SINR is calculated as shown in step 5 and eNBs with negative slope are discarded from set I (steps 6 and 7). To suppress impact of shadowing and fast fading, the slope is calculated over a whole period of prediction interval τ. For each VM placement, we select the eNB with the highest available capacity (step 10) and then eNB with the highest predicted gain in capacity is selected for VM placement (step 13). Following selection of eNB for VM placement, VM migration delay is predicted (step 14) and the option with lower delay between start of VM and VM migration is selected (step 15).

Algorithm 5 VM dynamic placement.

1: **for** $z \in Z$ **do**
2: **for** $i \in I$ **do**
3: predict $SINR_i(t, t+\Delta t, ..., t+\tau)$
4: predict $\tilde{c}_{z,i}(t, t+\Delta t, ..., t+\tau)$
5: $\alpha = \frac{dSINR_i}{dt}$
6: **if** $\alpha \leq 0$ **then**
7: $I = I \setminus i$
8: **end if**
9: **end for**
10: $\tilde{c}_z = \max_i\{\tilde{c}_{z,i}\}$
11: **end for**
12: $\hat{s}_{VM} = s_{VM}$
13: $s_{VM} = \arg\max_z(\tilde{c}_z - \tilde{c}_{current})$
14: $\tilde{t}_{\mathrm{M}} = \frac{G}{\tilde{c}_{\hat{s}_{VM}, s_{VM}}}$"
15: $option = min(t_{\mathrm{APP}}, \tilde{t}_{\mathrm{M}})$

Algorithm 6 PSwH with prediction.

1: **for** $i \in I$ **do**
2: predict $\tilde{c}_i(t, t+\Delta t, ..., t+\tau)$
3: **if** $s' = i$ **then**
4: $\rho_i = 1$
5: **else**
6: $\rho_i = 1 - t_{\mathrm{HO}}$
7: **end if**
8: **end for**
9: **while** $L > 0$ **do**
10: $q(t) = \arg\max_i(\tilde{c}_i(t) \times (\Delta t \cdot \rho))$
11: $L = L - max(\tilde{c}_i(t) \times (\Delta t \cdot \rho)$
12: **if** $i = q(t)$ **then**
13: $\rho_i = 1$
14: **else**
15: $\rho_i = 1 - t_{\mathrm{HO}}$
16: **end if**
17: $t = t + \Delta t$
18: **end while**

The enhancement of the PSwH by mobility prediction is described in Algorithm 6. First, available capacities of eNBs in set I are predicted (step 2) and handover vector $\rho = \{\rho_1, \rho_2, ..., \rho_{|I|}\}$ is initiated by setting its elements ρ_i to 1 if eNB i is also the serving eNB (step 4) or $1 - t_{\mathrm{HO}}$ otherwise (step 6). The handover vector is used for modification of communication capacity to each eNB as no data can be transferred between the UE and the eNB during t_{HO}. Until all required data L are transmitted (step 9), the eNB with the highest communication capacity is selected as $q(t)$. Note that the impact of handover on other services is discussed in [162]. Also, vector ρ is modified to be in line with $q(t)$ (step 12).

As the optimal solution for selecting VM placement and path selection is a combinatorial problem, it is required to go through every combination of serving eNB, VM

placement, and every step during offloading. This would have a large computation complexity and therefore, in our proposal, we reduce candidates for VM placement and path selection. Therefore, algorithm for VM placement has time-complexity of $O(|Z||I|\tau)$ and path selection $O(|I|\tau)$. Both time-complexities are lower than time-complexity of algorithm proposed in [130].

5.3.3 Performance Evaluation

In this section, models and scenario for performance evaluation are defined. The evaluation is carried out by means of simulations in MATLAB.

Simulation scenario and models

Major parameters of the simulation, presented in Table 5.4., are in line with recommendations for networks with small cells as defined by 3GPP in [157]. We also follow parameters of the physical layer and frame structure for LTE-A mobile networks defined in the same document.

Signal propagation is modeled according to 3GPP [157] with path loss model $PL = 128.1 + 37.6log10(d)$, where d is a distance between the UE and the eNB. A mapping function between SINR and MCS with BLER=10% is obtained from [176]. The offloaded task and results size is 200 kB [171], while each task contains $1e6$ instructions to be processed. The computation power of eNB and SCeNB is 3300 Millions Instructions Per Second (MIPS) [179]. The backhaul of eNBs is modeled as optical fiber with capacities (in Mbit/s) generated from normal distribution with $\mu = 100$ and $\sigma^2 = 2$.

Table 5.4. Simulation parameters

Parameter	**Value**
Simulation area	800 x 800 m
Carrier frequency	2 GHz
Bandwidth for downlink/uplink	10/10 MHz
Tx power of eNB/SCeNB/UE	27/15/10 dB
Number of eNB/SCeNB	19/57
VM size/start time	20 MB/500 ms
Offloaded task/results size	200/200 kB
Offloaded task number of instructions	$1e6$ instructions
eNB/SCeNB CPU	3300 MIPS
Prediction window τ/Wang's algorithm	20s/60s
Prediction accuracy	90%
Shadowing factor	6 dB
Handover interruption duration	30 ms
Number of UEs	200
Speed of users	1 m/s
Backhaul capacity-Normal distribution	$\mu = 100, \rho^2 = 2$ Mbit/s
Simulation time T	2 000 s
Number of simulation drops	10 drops

Since we target to real-time applications, offloaded task has a size of 200 kB (as in [180] authors consider task to be in tenths of kB) and its arrival rate is specified by λ. The size of data transferred during VM migration (data in RAM of offloaded task) is 20 MB. Time before VM is prepared to process an offloaded task (start time) is 500 ms, which consist only of starting an offloaded application at the VM. Radio and backhaul resource allocation is done by round-robin scheduling.

We assume the hexagonal grid of 19 eNB like in [130] and we further drop 57 HeNBs into the simulation area. There are 200 UEs moving within the area of all eNBs according to smooth random mobility model. The prediction accuracy of users' mobility in simulations is based on percentage of correct predictions in [181], i.e., it is 90% in our case.

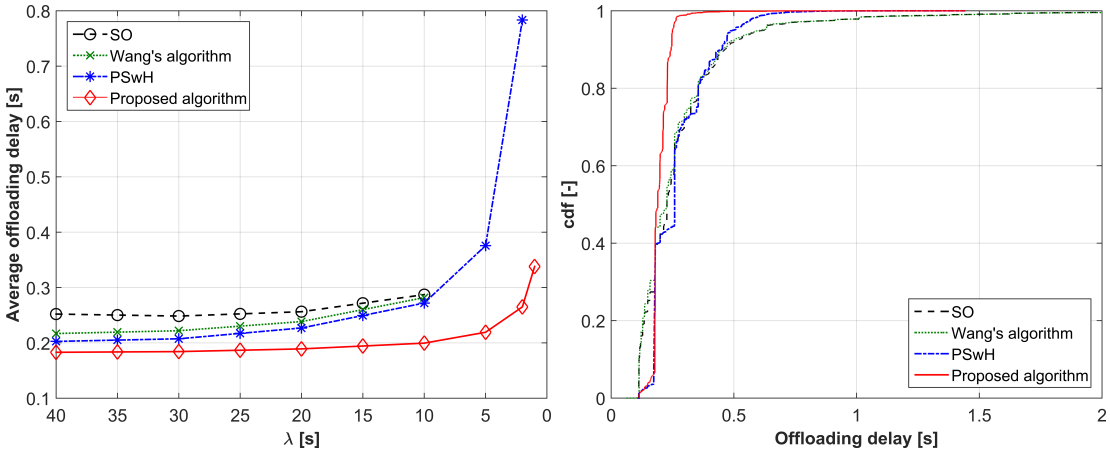
Performance evaluation

In our simulations, the proposed algorithm is compared with three competitive algorithms:

- SO [175] - The VM is kept at the serving eNB, so the VM is migrated each time handover is performed.
- Wang's algorithm [130] - VM placement is based on predicted future costs of its placement.
- PSwH [162] - Communication path (serving eNB) is selected so to minimize communication delay.

In Figure 5.13a we show the average offloading delay (consisting of uploading offloaded task, computing, and collecting results) of the task in dependency on task inter-arrival rate (λ). From this figure, we can see that with decreasing λ, the average offloading delay increases as the load of communication and computation resources increases. The proposed algorithm reduces the average offloading delay significantly comparing to all competitive algorithms. For lightly loaded network ($\lambda = 40s$), the average offloading delay is reduced by the proposed algorithm by 27.3%, 15.6%, and 9.7% with respect to the SO, Wang's algorithm, and PSwH, respectively. For heavily loaded network ($\lambda = 10s$) the gain is 30.5%, 29.2%, and 26.6% with respect to the SO, Wang's algorithm, and PSwH, respectively. The gain is caused by cooperation between VM placement and path selection according to predicted situation in the network.

Note that results for the SO and Wang's algorithm for $\lambda < 10s$ are not depicted as these algorithms cannot handle such load of network as delay of tasks can lead to tasks being buffered at the UE and thus leading to congestion of communication and computation resources. The proposal, by combining both VM placement and path selection avoids over utilized eNBs a thus works even for $\lambda = 1s$. The proposal outperforms all compared algorithms as compared to the PSwH, which has the second lowest delay, reduces offloading delay by up to 66%.



(a) (b)

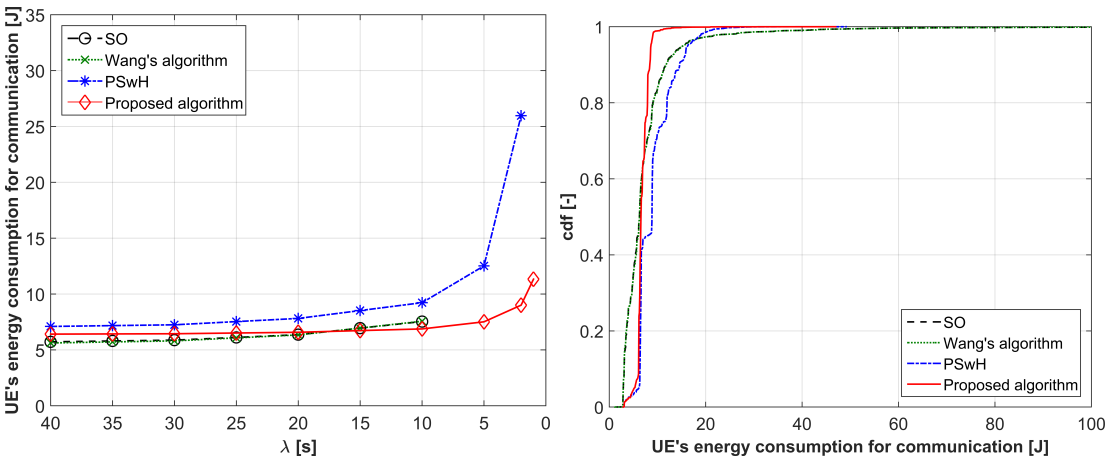
Figure 5.13. Offloading times required to offload, compute and collect results of the offloaded task, Average time (a) and CDF of time (b).

In Figure 5.13b, we compare CDF of the average offloading delay for $\lambda = 10s$. We show CDF for $\lambda = 10s$ as it corresponds to heavily loaded network, which is more challenging than lightly loaded network.

The offloading delays reached by UEs in case of the SO, Wang's algorithm, and PSwH are spread significantly from relatively low values (115 ms) to extremely high delays not acceptable for real-time services (even more than 2s). Contrary, the proposed algorithm offers stable delay around 200 ms for almost all UEs. For example, the delay experienced by 95% of UEs is below 250ms for the proposal while competitive SO, Wang's algorithm, and PSwH requires 610ms (144%more), 610ms (144% more), 500ms (100% more), respectively. Consequently, almost all UEs exploiting the proposed algorithm can exploit real-time services with high quality.

In Figure 5.14a, comparison of average energy consumed by the UE for communication of a single task is shown. With decreasing λ, consumed energy increases as offloading delay is higher due to increased network load and relation between UE's energy consumption and delay [149]. From Figure 5.14a, we can see, that the PSwH is the most energy hungry and it consumes between 10.7% and 188% more energy than the proposed algorithm. The SO and Wang's algorithm require less energy (up 9%) per offloaded task than the proposed algorithm if the network is lightly loaded ($\lambda > 15s$). Contrary, for heavily loaded network ($\lambda < 15s$), the proposed algorithm becomes more energy efficient (saving of 9%). The reason for increase in energy consumption by the proposal at light network load is the fact that the proposed algorithm targets solely on offloading delay and disregard energy consumption. Extension towards consideration of the energy consumption is considered as a future work. Note also that the SO and Wang's algorithm cannot serve tasks with λ lower then 10s.

In Figure 5.14b, we show CDF of the energy spent by the UE for communication for $\lambda = 10s$. The energy consumption reached by the SO and Wang's algorithm is spread more wide so energy consumption of some UEs is reduced comparing to the proposed



(a) (b)

Figure 5.14. Energy consumption of UE communication, Average energy (a) and CDF of energy consumption (b).

algorithm while some UEs consumes significantly more energy. This shows fairness of the proposed algorithm among users and there are no users significantly punished for unfair allocation of resources for computation. The energy consumed by 95% UEs is below 8.61J for the proposal while competitive SO, Wang's algorithm, and PSwH consumes 15.3J (77.7%more), 15.3J (77.7% more), 16.3J (89.3% more), respectively.

5.3.4 Conclusion

In this section, we have proposed an algorithm for dynamic allocation of computing and communication resources for the MEC. The algorithm dynamically places VMs considering load of eNBs and selects communication path between the UE and the eNB with allocated VM. The algorithm is based on MDP and exploits mobility prediction with a known prediction accuracy.

Comparing to state of the art approaches, the proposed algorithm reduces the offloading by 10-66%. The superiority of the proposed algorithm is more notable for high arrival rate of the offloading requests, i.e., for heavily loaded network. At the same time, the energy consumed by the UEs for offloading is kept at similar level as for the state of the art algorithms. The proposed algorithm also balances fairness among users in terms of experienced delay and energy consumption so that all UEs can exploit real-time services even for very high arrival rates of the offloading requests.

5.4 Prediction based communication and computing resource allocation for MEC

The joint communication and computation resource allocation in the previous section is limited to a scenarios with a known and fixed mobility prediction accuracy. Therefore, to overcome this limitation, we present a solution consisting of the UE's mobility and channel quality prediction framework with an algorithm for joint communication and

computation resource allocation, that enable exploitation of the MEC for a real-time offloading.

The contribution and novelty of the proposed solution in this section is summarized as follows:

- We propose a novel algorithm for the DCCRA for MEC systems. Unlike the existing works, we target the offloading of the real-time applications by the moving UEs. This implies requirements on a very low delay. We solve this problem via two cooperating sub-algorithms, one for the dynamic selection of the communication path (i.e., the gNB that serves the UE) and one for the VM placement.
- To facilitate the proposed DCCRA, we develop a framework for a prediction of the UE's mobility and channel quality based on a probabilistic model of the UE's mobility. The mobility prediction is first illustrated in a scenario with one degree of mobility freedom and, then, we generalize it for multiple degrees of mobility freedom.
- Via simulations, we show that the proposed algorithm enables offloading of the real-time applications by the mobile UEs and keeps the offloading delay under 100 ms even for a high arrival rate of up to five tasks per second per UE. Such performance is notably superior to existing works and it facilitates exploitation of the MEC services by the real-time applications even for the moving UEs. Furthermore, we show that the performance of the proposed solution in terms of the offloading delay and the energy consumption of the UEs is significantly improved comparing to existing solutions and it is even close to the case with a perfect prediction of the channel quality.

The rest of this section is organized as follows. In the next section, the resource allocation problem is formulated and assumptions along with a system model are described. In Section 5.4.2, the framework for the mobility and channel prediction suitable for the proposed resource allocation is outlined. The proposed resource allocation algorithm is defined in Section 5.4.3. Then, in Section 5.4.4, the environment and models for simulations are presented, and the simulation results are discussed in Section 5.4.5. Last, Section 5.4.6 concludes this section and the proposed solution.

5.4.1 System Model and Problem Formulation

In this section, we define the system model exploited for the proposed algorithm, we formulate the computing and communication resource allocation problem for the offloading, and we summarize the main assumptions for the resource allocation algorithm.

System Model

We consider a set $S = \{s^1, s^2, \ldots, s^M\}$ of the gNBs and a set $U = \{u^1, u^2, \ldots, u^N\}$ of the UEs. The serving gNB for the UE at the discrete time t, denoted $s_t \in S$, is selected

as the gNB providing the highest RSS. As the UE moves, the serving gNB is updated following a conventional hard handover procedure based on the RSS considering also a handover interruption with a duration of t_{HO}. This means that the serving gNB is updated if there exists the gNB $s' \in S$, where $s' \neq s_t$, for which $RSS(s') > RSS(s_t) + \Delta_{HO}$, where Δ_{HO} is the handover hysteresis (see, e.g., [182] for more details about the conventional hard handover and hysteresis in mobile networks). Based on the serving gNBs determined for the UEs, we define $n_t^R(s)$ as the number of UEs sharing the radio communication resources of the s-th gNB at the time t, respectively. Furthermore, in a similar way, we define $n_t^B(s)$ as the number of UEs sharing the backhaul communication resources of the s-th gNB at the time t, respectively.

Then, we define $s_t^* \in S$ as the gNB where the VM or the container for the UE is placed at the time t. We assume the possibility to pre-allocate the VMs or the containers on multiple gNBs to alleviate the issue of an unreliable mobility and channel predictions. Nevertheless, only one VM or container is exploited by the application offloaded by each UE at any given time. The time required to start the VM, including the VM pre-allocation, on the gNB is denoted as t_{VM}. Next, we define $\omega_t(s)$ as the amount of available processing resources of the s-th gNB in MIPS at the time t. Then, the MIPS requirements of the application offloaded by the u-th UE is labeled as $\omega(u)$. The s-th gNB is considered for the VM placement of the u-th UE at the time t only if the following condition holds:

$$\omega_t(s) > \omega(u). \tag{5.11}$$

Note, that the offloaded application requires not only the computing power, but also memory and/or hard drive capacity. These resources can be formulated in the same way as for the computing power requirements and an extension of the condition (5.11) to these parameters is straightforward. Therefore, without loss of generality and in order to keep our notation simple, we consider (5.11) as the only resource restriction of our problem. The MEC system model with the gNB communication and computing load is shown in Figure 5.15.

The offloaded task is defined by the amount L_{O} of offloaded data (in bits), the amount L_{C} of collected data representing the computation results (again in bits), and the number L_{P} of instructions of the offloaded task to be executed at an gNB. The offloaded tasks are generated by the offloaded application with a task arrival rate λ, representing the number of offloaded tasks generated per second. The offloaded task can be delivered from the UE to the computing gNB directly via radio if $s_t = s_t^*$ or indirectly via the serving gNB s_t, if the serving gNB is different from the computing gNB s_t^*, i.e., if $s_t \neq s_t^*$. The latter case can appear, for example, in the situation when the serving gNB is not able to offer a sufficient computing power to the UE. The latter case assumes to exploit the backhaul connections of the s_t^* and the s_t for a transfer of the offloaded task between the serving and the computing gNBs. Note that the backhaul communication between the s_t and the s_t^* is assumed to be routed via an operator's core network as it is done in conventional mobile networks [183, 184].

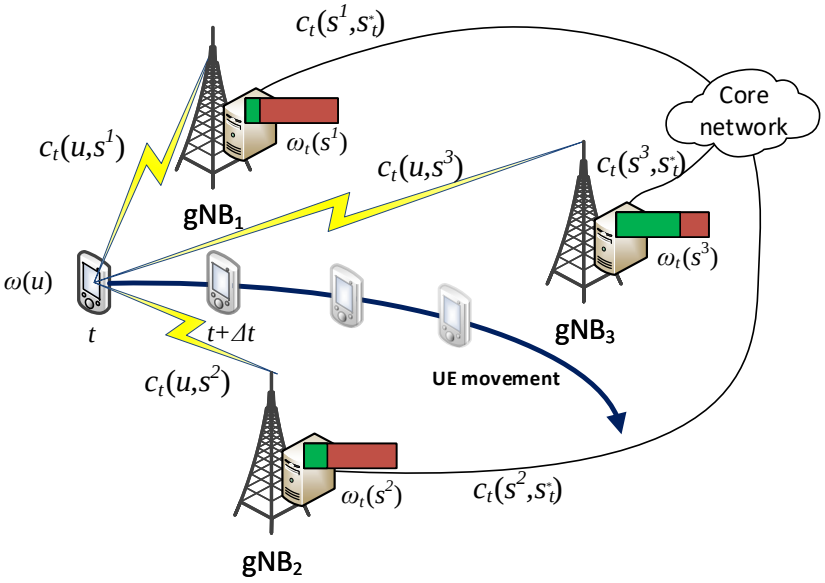


Figure 5.15. MEC system model with one mobile UE.

Furthermore, we define the set $Q_t(u) \subseteq S$ as the subset of all gNBs with which the u-th UE can communicate at the time t. In particular, this set contains only the gNBs to which the UE has SINR above a minimumSINR level required for communication ($SINR_{min}$). For the communication between the UE and the serving gNB, we assume LTE or 5G-based radio interface with radio resources shared equally among all UEs connected to the same gNB. Thus, the radio communication data rate between the u-th UE and the s-th gNB is calculated as:

$$c_t(u,s) = \nu_u \rho_u \frac{n^{RB}(s)}{n_t^R(s)}, \tag{5.12}$$

where ν_u is the number of bits per symbol for a given modulation scheme, ρ_u is the code rate used for the radio communication between the gNB and the u-th UE, and $n^{RB}(s)$ is the number of RBs available at the s-th gNB. Both ν_u and ρ_u are derived from the channel quality according to the SINR of the u-th UE (see [176] for more details).

When $s_t \neq s_t^*$, the data rate expected on the backhaul connection between the serving gNB s_t and the computing gNB s_t^* (see Figure 5.15) is defined as:

$$c_t(s,s^*) = min\left\{\frac{c_s}{n_t^B(s)}, \frac{c_{s^*}}{n_t^B(s^*)}\right\}, \tag{5.13}$$

where c_s and c_{s^*} denote the available backhaul capacity of the serving gNB s_t and the computing gNB s_t^*, respectively. Note that the "min" in (5.13) indicates that different data rates can be expected on the backhauls belonging to s_t and s_t^*.

The communication data rate available for a delivery of the offloaded task from the UE to the s_t^* either directly via radio (if $s_t = s_t^*$) or indirectly via the radio of s_t and the

backhauls of s_t and s_t^* (if $s_t \neq s_t^*$) is derived as:

$$c_t^{UL}(u, s, s^*) = \begin{cases} c_t(u, s) & \text{if } s_t = s_t^* \\ min\{c_t(u, s), c_t(s, s^*)\} & \text{otherwise} \end{cases}, \tag{5.14}$$

Problem Formulation

Our objective is to find an allocation strategy of the computing and communication resources that minimizes the total offloading metric, represented by the offloading delay. Thus, the objective is to find the resource allocation strategy that minimizes the total offloading delay for the UE (denoted as t_{MEC}). Minimization of the total offloading delay enables offloading of the real-time tasks, as these tasks require a very low delay. The total offloading delay consists of:

i) the time required to deliver the offloaded task from the UE to the gNB that starts the computation, determined as:

$$t_{\text{O}} = \frac{L_{\text{O}}}{c_t^{UL}(u, s, s^*)}, \tag{5.15}$$

ii) the time required to process the offloaded task, calculated as:

$$t_{\text{P}} = \frac{L_{\text{P}}}{\omega_t(s)}, \tag{5.16}$$

iii) the time required to deliver the processed data from the gNB that finishes the computation to the UE, defined as:

$$t_{\text{C}} = \frac{L_{\text{C}}}{c_t^{DL}(u, s, s^*)}, \tag{5.17}$$

with data rate $c_t^{DL}(u, s, s^*)$, derived in line with (5.14), but for the downlink, as the results of computation are received by the UE,

iv) the time consumed by the handover process, defined as:

$$t_{\text{H}} = \sum_{i=1}^{n^{\text{H}}} t_{\text{HO}}^i \tag{5.18}$$

where t_{HO}^i is the duration of the i-th handover and n^{H} is the number of handovers due to the UE changing serving gNB.

v) the time of the VM starts (including obtaining UE's application which processes offloaded tasks) during the offloading, determined as:

$$t_{\text{M}} = n^{\text{VM}} t_{\text{VM}} \tag{5.19}$$

where n^{VM} is the number of VM starts (equal to 0 if no VM starts is needed or VMs are

pre-allocated) taking place during the offloading process. Note that the gNB that receives the offloaded data and the gNB that delivers the results back to the UE may not be the same due to the UE's mobility. The total offloading delay experienced by the UE is then calculated as:

$$t_{\mathrm{MEC}} = t_{\mathrm{O}} + t_{\mathrm{P}} + t_{\mathrm{C}} + t_{\mathrm{M}} + t_{\mathrm{H}}. \tag{5.20}$$

We can treat the minimization problem as a pair of joint problems. The first problem is the determination of the sequence of gNBs $\{s_t^*\}_t^{\mathrm{opt}}$, where the VM (or the container) should be placed for the u-th UE at each time t. The second problem is the selection of the communication path, identified with the serving gNBs sequence $\{s_t\}_t^{\mathrm{opt}}$. Merging both problems, we formulate the objective as:

$$\{s_t^*\}_t^{\mathrm{opt}}, \{s_t\}_t^{\mathrm{opt}} = \underset{\{s_t^* \in S\}_t, \{s_t \in S\}_t}{\arg\min} t_{\mathrm{MEC}}. \tag{5.21}$$

However, solving (5.21) is, in general, difficult and impractical as both computing and communication resource allocation have to be done together for each t leading to a complex problem. Furthermore, the offloading delay (5.20), consisting of the time to offload the task (5.15) and the time to collect the processed results (5.17), depend on the data rates defined in (5.14). These data rates are not always known and should be predicted. This complicates the possibility to reach the global optimum. Moreover, finding the global optimum at each t leads to allocation of the computing resources (VMs) at different gNBs due to variation of the channel quality over time. Exploiting the VMs on different gNBs then leads to a high number of VM starts (t_M) and handovers (t_H) as shown in [185]. Thus, even though the global optimum is known, it may be impossible to reach it in practice, as the VMs would be constantly started over and over again.

Since the problem (5.21) is impractical and cannot be directly solved, we simplify the problem and transform it into the maximization of the communication data rate $c_t^{UL}(u, s, s^*)$ due to the constant L_{O} and L_{C}, while considering (5.11), (5.13), (5.14), and (5.20). We focus on the uplink communication rate, because the uplink is commonly assumed to be of a lower data rate than the downlink. The extension to consider the downlink communication rate is trivial and we leave it out to simplify the notations. Therefore, we transform the problem into the following:

$$\{s_t^*\}_t^{\mathrm{opt}}, \{s_t\}_t^{\mathrm{opt}} = \underset{\{s_t^* \in S\}_t \{s_t \in S\}_t}{\arg\max} \{c_t(u, s, s^*)\}_t \tag{5.22}$$

$$\text{s.t.} \quad \omega_t(s^*) > \omega(u). \tag{5.23}$$

where the constraint (5.23) is defined to avoid placing the VMs on the gNBs, which do not have enough computing power to host the VM for the UE. The constraint (5.23) considers computing power of the gNBs, since each gNB can have different computing power. Since $\{s_t^*\}_t^{\mathrm{opt}}$ and $\{s_t\}_t^{\mathrm{opt}}$ can be different, the transformed problem can be solved as two subproblems via two proposed cooperative algorithms.

Assumptions

In this section, we assume that every task is offloaded, as assumed in [130]. The assumption of offloading every task represents the case of the UE that does not have enough computing resources to process tasks itself and is forced to offload them. Note that introducing the offloading decision simply leads to a lower amount of tasks to be processed in the MEC servers, as only some tasks would be offloaded. Thus, the proposed solution is applicable to any offloading decision algorithms and its impact on performance is proportional to changes in the task arrival rate λ investigated later in this section.

In [130] and [185], the authors suppose that the communication data rate is predicted with a pre-defined fixed accuracy, but this assumption is quite strong. More realistically, here, we assume that the prediction accuracy is unknown and varies in time, or even that a prediction is unavailable at all. This reflects the unreliability of the UEs' mobility prediction strategies, even when they are based on a significant amount of information about the UEs [186]. A suitable approach is to exploit probabilistic models or probabilistic-free models as in [187]. To design and implement such approach, we assume that the knowledge of users' contextual information, such as scheduled meetings, favorite places, etc. as exploited, e.g., in [186, 187] is *not* available. Such assumption complicates the prediction and potentially negatively impacts on the performance of the developed algorithm. However, this assumption is motivated by questionable willingness of the users to provide this type of information to the network operator due to privacy issues. Thus, we expect the availability of only the information that is typically available to the network or which is commonly shared by the users. More specifically, we exploit anonymized UEs positions and SINR at those positions. As this type of information can be easily anonymized, the privacy risks are significantly lowered with respect to [186, 187].

For clarity and simplification of explanation, we assume the architecture where the MEC servers are collocated with the gNBs as proposed in [19, 101]. Note that a placement of the MEC servers to other network nodes, such as core network elements, increases t_{O} and t_{C}. Since the proposed algorithm considers t_{O} and t_{C} in terms of the communication data rate, it can be simply extended to consider also different placements of the MEC servers. However, this extension is omitted here for the sake of clarity.

The data of the offloaded task are processed via an UE application in the MEC. For this, we assume, that the UE's application at the MEC server (represented by a gNB or Small Cell gNB (SCgNB)) is obtained from a cloud storage during the VM start time.

5.4.2 Mobility and Channel Quality Prediction

To solve the problem formulated in the previous section, we develop the mobility and channel quality predictions with a low complexity to derive the expected communication data rate. The predicted data rate is then exploited by the proposed computing and communication resource allocation algorithm. The effectiveness of different mobility prediction approaches depends on the application scenarios. Therefore, we split the

description of the mobility prediction into two cases with: i) one degree of mobility freedom (e.g., movement along a sidewalk or street with no possibility to turn away), and ii) multiple degrees of movement freedom with possibility to change the direction (e.g., crossroads, open spaces, squares, etc.). These two cases are explained in the next two subsections. Then, the last part of this section describes the proposed channel quality prediction strategy, which is further exploited for the communication data rate prediction.

Mobility prediction with one degree of mobility freedom

If the UE's mobility is limited to one degree of mobility freedom (i.e., UE following a sidewalk or street) an extrapolation of the UE's movement is a suitable approach to predict the UE's future position following the assumption of the limited knowledge about the UEs as defined in Section 5.4.1. The prediction of the UE's movement can be divided into two subcases: i) the UE moving along a straight path and ii) the UE moving along a curved path. In the first subcase, we can simply extrapolate the future movement of the UE from its past movement. However, in the second case, a linear motion extrapolation of the UE's movement would lead to an inaccurate prediction of its position, crossing the environment boundaries such as sidewalks, streets or walls, as shown in Figure 5.16. Therefore, after describing the extrapolation of the UE's movement, we also outline how to exploit the knowledge of the environment to obtain a more accurate prediction of the UE's movement.

The position of the UE at the discrete time t is represented by the coordinates (x_t, y_t). From the current time instant t and the previous time instant $t - \Delta t$, we obtain the UE's approximated velocity vector $(\Delta x, \Delta y)$ where:

$$\Delta x = \frac{x_t - x_{(t-\Delta t)}}{\Delta t}, \tag{5.24}$$

$$\Delta y = \frac{y_t - y_{(t-\Delta t)}}{\Delta t}. \tag{5.25}$$

The predicted UE's position at the time $t + k\Delta t$, where $k = \{1, 2, \ldots, K\}$ and $K\Delta t$ being the prediction window in seconds (typically ranging up to tens of seconds [43]), is calculated as:

$$x_{t+k\Delta t} = x_t + k\Delta x\Delta t, \tag{5.26}$$

$$y_{t+k\Delta t} = y_t + k\Delta y\Delta t. \tag{5.27}$$

Now we describe an extension of the simple linear extrapolation defined in (5.26) and (5.27) by exploiting a knowledge of the environment. Let the UE be located at the position (x_t, y_t) (indicated by a dot in Figure 5.16) and let the UE follow a curved street as shown in Figure 5.16. In our model, the street is represented by a discrete set of street centers $X \times Y = \left\{(x_{(j)}, y_{(j)})\right\}_{j \in J}, J \subseteq \mathbb{Z}$, indicated by the crosses in Figure 5.16. To exploit the knowledge of the environment, the UE's position is mapped to the closest street center,

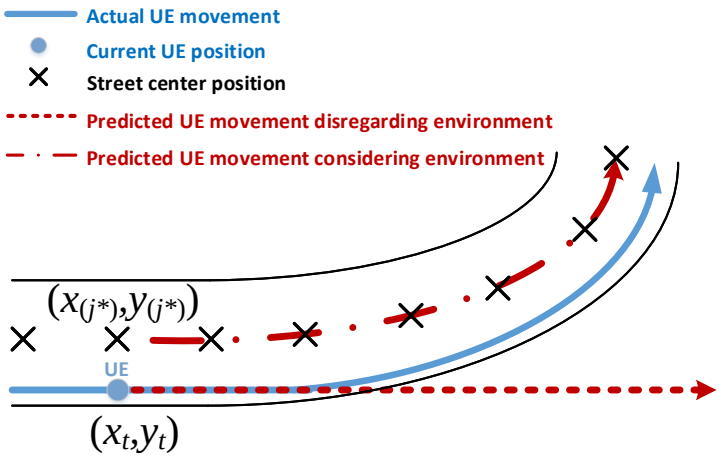


Figure 5.16. Example of UE mobility prediction with one degree of mobility freedom following a curved street with known street center positions.

identified by the index j^* determined as follows:

$$j^* = \arg\min_{j \in J} \sqrt{\left(x_t - x_{(j)}\right)^2 + \left(y_t - y_{(j)}\right)^2}. \tag{5.28}$$

In Figure 5.16, the closest street center to the UE is $\left(x_{(j^*)}, y_{(j^*)}\right)$. Based on the knowledge of the environment, the UE's position at $t + k\Delta t$ is then mapped to the street center indexed by $j^* + \kappa(k)$ as:

$$x_{t+k\Delta t} = x_{(j^*+\kappa(k))}, \tag{5.29}$$

$$y_{t+k\Delta t} = y_{(j^*+\kappa(k))}. \tag{5.30}$$

where $\kappa(k) = \left\lfloor k\Delta t \frac{\sqrt{\Delta x^2 + \Delta y^2}}{\Delta j} \right\rfloor$ approximates the number of street centers run over by the UE during k time instants and Δj is the distance between any two consecutive street centers, which we consider constant and can be computed as

$$\Delta j = \sqrt{\left(x_{(j+1)} - x_{(j)}\right)^2 + \left(y_{(j+1)} - y_{(j)}\right)^2}. \tag{5.31}$$

Mobility prediction with multiple degrees of movement freedom

Now, we extend our mathematical formulation to the case where the UE has multiple degrees of mobility freedom. The set of degrees of freedom for the UE movement is denoted as W. This set includes the angles w that the UE can select for its future direction. The set of arrival angles V includes the angles v from which the UE has arrived to the current position. In a general scenario, the UE can select any departure (arrival) angle between 0° and 360°. To limit the complexity, we discretize the angles in a similar way as in [187]. This means that a range of nearby angles is represented by a single departure angle. For

example, the discretization with angle difference of 1^{o} results in 360 elements (arrival and departure angles) in the both sets V and W. An example of the UE with four degrees of freedom, i.e., $|V| = |W| = 4$, is shown in Figure 5.17, where the UE arrives from the angle v_3 and can depart in the direction of any angle from the set $\{w_1, w_2, w_3, w_4\}$.

In our model, among all the departure angles in W, we consider only those with non-zero probability of selection. The adopted probabilistic model is based on Markov chains with underlying Hidden Markov Model (HMM), which is suitable for systems with multiple states and transitions between those states. The important property of Markov chains is that the conditional probability distribution of future states depends solely on the present state. This property is valid for our model as we can legitimately suppose that the departure angle selected by the UE depends only on the arrival angle v and the current UE's position.

The HMM model consists of states and transition probabilities between the states. The states of the HMM model (represented by departure angles) are learned from an environment layout in the form of a map, (e.g., openstreetmaps.org [188]). Exploitation of the environment maps for mobility prediction is considered for example in [186] or [187]. Therefore, with the known states, only the transition probabilities (probability of the transition from the arrival to departure angles) of the Markov chain need to be estimated. The transition probabilities represent the probability that each departure angle is chosen [189]. Estimation of the transition probabilities is then done by estimation of the transition probabilities of the Markov chain as described in [189]. To this end, the number of transitions from each arrival angle $v \in V$ to each departure angle $w \in W$ is counted. Note that, if the number of states in unknown, the estimation of the HMM states and transition probabilities is done via the Maximum Likelihood Estimation (MLE) [190].

As the time to learn the transition probabilities between each arrival and departure angle can be high, we consider the transition model aggregated over all the UEs altogether, which reduces the learning time for the estimation of the transition probabilities between the states in the Markov chain. The cost of this aggregation is a slightly lower accuracy of the learned model. However, once enough transitions for each UE are collected, the transition model of the individual UE can be used to replace the aggregated model. It is worth to mention that the learned aggregated transition model still guarantees good results, because the main purpose of the model is to avoid the transitions with very low probabilities. Moreover, by exploiting the aggregated transition model, the transition probability of any UE, including those with unknown transition model, can be predicted.

The probability that the UE at the position (x_t, y_t) selects a departure angle w conditioned by the arrival angle v is denoted $P\left(w|v, (x_t, y_t)\right)$. This probability, representing the transition probability in the Markov chain [189], is calculated as:

$$P\left(w|v, (x_t, y_t)\right) = \frac{N\left(v, w, (x_t, y_t)\right)}{\sum_{w' \in W} N\left(v, w', (x_t, y_t)\right)}, \tag{5.32}$$

where $w \in W$ is the selected departure angle, $N\left(v, w, (x_t, y_t)\right)$ is the number of transitions from the arrival angle v to the departure angle w at the UE's position (x_t, y_t) summed

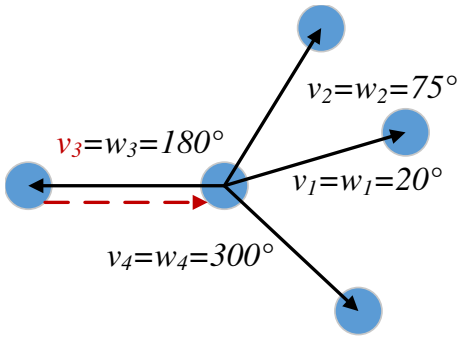


Figure 5.17. Example of the UE with multiple degrees of mobility freedom, arriving from angle v_3 (red dashed line) with multiple options for the departure angle w (solid lines).

up till the current time t and for all UEs. Notice that we do not exclude the possibility that the UE stops at the crossroad or departs via the arrival angle (i.e., $w = v$). In this case, the VM placement remains constant, because frequent re-deployments or migrations would overload the network and lead to a disruption in the MEC service. However, the communication path selection is exploited to provide sufficient connectivity considering also channel changes.

Based on the probabilistic model (5.32), we extend the prediction of mobility by considering the departure angles w with non-zero probability. When, the surrounding environment is unknown, for a given w, (5.26) and (5.27) are modified as follows:

$$x^w_{t+k\Delta t} = x_t + k\Delta t\sqrt{\Delta x^2 + \Delta y^2}\cos(w), \tag{5.33}$$

$$y^w_{t+k\Delta t} = y_t + k\Delta t\sqrt{\Delta x^2 + \Delta y^2}\sin(w), \tag{5.34}$$

where Δx and Δy are calculated via (5.24) and (5.25), respectively. Furthermore, in the case when the movement is predicted with environment knowledge, we extend (5.28) to:

$$j^*_w = \underset{j\in J_w}{\arg\min}\sqrt{\left(x^w_{t+\Delta t} - x_{(j)}\right)^2 + \left(y^w_{t+\Delta t} - y_{(j)}\right)^2}, \tag{5.35}$$

where J_w is the set of street centers along the departure angle w. Note that j^*_w represents the closest street center to the UE's position at the time $t+\Delta t$, when the departure angle is w. Accordingly, (5.29) and (5.30) are generalized as follows:

$$x^w_{t+k\Delta t} = x_{(j^*_w+\kappa(k-1))}, \tag{5.36}$$

$$y^w_{t+k\Delta t} = y_{(j^*_w+\kappa(k-1))}, \tag{5.37}$$

with $\kappa(0) = 0$.

From the estimated positions of the UE at the times $t + k\Delta t$, we calculate the corresponding Euclidean distances to the gNBs. These distances replace the communication data rate as the offloading metric whenever the data rate is unknown or impossible to predict. The predicted Euclidean distance between the UE and the s-th gNB located at

$(x(s), y(s))$ at the time $t + k\Delta t$ is calculated as:

$$d_{t+k\Delta t}(s, w) = \sqrt{\left(x^w_{t+k\Delta t} - x(s)\right)^2 + \left(y^w_{t+k\Delta t} - y(s)\right)^2}. \tag{5.38}$$

SINR and communication data rate prediction

After predicting the UEs' future movement, the communication data rate is calculated based on the estimated future SINR values. Future SINR is predicted either from SINR maps [94] or is extrapolated from the past SINR values if the SINR map is not learned yet. First, we describe the exploitation of the SINR map, and then we describe the extrapolation of theSINR based on the past SINR values.

The SINR map, shared by all the gNBs is represented by a matrix Ψ containing the SINR levels $\Psi_{x,y}$ observed by the UEs at discrete and quantized coordinates $x \in \mathbb{N}$ and $y \in \mathbb{N}$. The SINR map is updated each time when the SINR measurement is received from the UE at the coordinates (x, y) and stored in $\psi_{x,y}$. The update of the SINR map is implemented as a weighted average of the current SINR map value $\Psi_{x,y}$ and $\psi_{x,y}$. Then, the SINR map is updated as follows:

$$\Psi_{x,y} \leftarrow \left((1 - \chi)\, \Psi_{x,y} + \chi \psi_{x,y}\right), \tag{5.39}$$

where χ is the weight of the new input value to the SINR map. Note that χ can be optimized based on the performance in a real deployment. Due to the dependency of χ on the real deployment, we leave the optimization of χ for future research.

If the SINR map is not learned yet, the SINR is extrapolated based on Autoregressive integrated moving average (ARIMA) [191], because it enables prediction of non-stationary SINR as required in our case. The SINR is non-stationary due to its time variance caused by varying power levels of received and interference signals.

The generic ARIMA (P, D, G) model is defined by the order of autoregressive part P, the degree of the first differencing D, the order of the moving average part G, and the model parameters: the autoregression θ_i and the differencing and moving average terms ϕ_i (index i indicates terms of autoregression and terms of moving average). As the SINR does not periodically change values, we leave out the seasoning difference, which is a common part of the generic ARIMA, but it is exploited only if the predicted time series depends on the month, hour, and so on. For our purposes of the SINR level prediction,

we define the ARIMA model for SINR prediction as:

$$B^P SINR_t = SINR_{t-P}, \tag{5.40}$$

$$\phi_i(B) = 1 - \sum_{i=1}^{G} \phi_i B^i, \tag{5.41}$$

$$\theta_i(B) = 1 - \sum_{i=1}^{P} \theta_i B^i, \tag{5.42}$$

$$SINR_t = \frac{\theta_G\,(B)\, e_t}{\phi_P\,(1-B)^D}, \tag{5.43}$$

where $SINR_t$ is the SINR time series, B^i is the lag operator of the i-th order, and e_t is the error term of the ARIMA model.

The ARIMA model and the coefficients of autoregression, moving average, and lag operator are estimated from the past samples of SINR by MLE following [191]. Then, the future SINR levels $SINR_{t+\Delta t}$, $SINR_{t+2\Delta t}$, $\ldots$, $SINR_{t+K\Delta t}$ are calculated based on the estimated ARIMA model and the coefficients from (5.43). The communication data rate is then predicted from SINR levels at times $(t+\Delta t, t+2\Delta t, \ldots, t+K\Delta t)$ via (5.12). Note that K represents the number of predicted SINR samples.

5.4.3 Proposed Dynamic Communication and Computing Resource Allocation Algorithm

In our previous work [185], we have shown that if a prediction with a fixed accuracy is available and the VMs are pre-allocated on all gNBs, the communication and computing resource allocation can handle offloading of the tasks with the arrival rate λ up to 1 task per second in the considered scenario. However, the hypothesis of fixed prediction accuracy is not reasonable for real networks. Thus, we propose an algorithm, denoted DCCRA, which exploits the probabilistic UEs' mobility prediction approach described in Section 5.4.2.

The DCCRA is composed of two cooperating algorithms: one for the computing and one for the communication resource allocation. The computing part targets a proper VM placement (computing resource allocation) while the communication part consists in selection of a proper communication path (communication resource allocation). The cooperation of the proposed algorithms is shown in Figure 5.18 for a prediction window of $K\Delta t$. First, the VM placement is determined for each UE via Algorithm 7 over the duration of $K\Delta t$. Then, a proper communication path is selected by Algorithm 8 for individual UEs in every time interval t. Both parts of the DCCRA are described in the following subsections, respectively, followed by a complexity analysis.

Note that the computing resources can be allocated either in the form of the VMs or the containers. To simplify the following text, we describe the algorithm for the VMs, but these are interchangeable with the containers in the proposed algorithm. Furthermore,

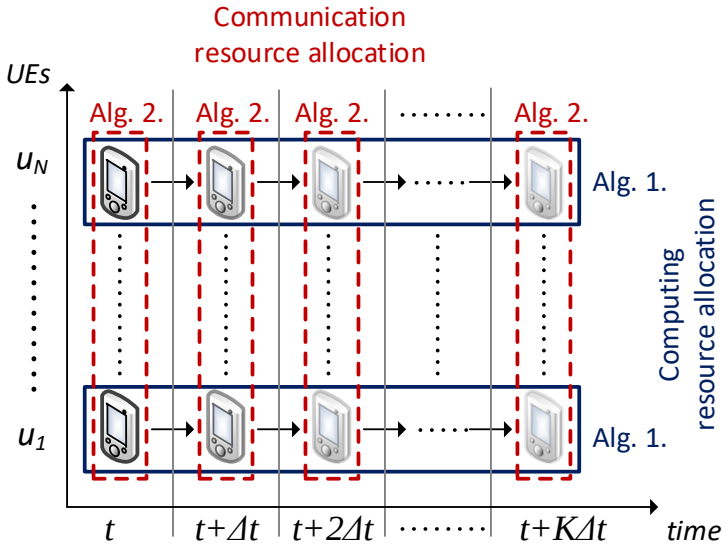


Figure 5.18. Cooperation of the proposed DCCRA algorithms.

the proposed algorithm is designed for a generic case with multiple degrees of mobility freedom, as described in Section 5.4.2.

Computing resource allocation

The computing resource allocation part of the proposed DCCRA algorithm, deciding where and when to allocate the VMs for each UE, is described in this sub-section.

In general, the DCCRA selects the most suitable gNB s_t^* for the placement of computing resources in terms of the VMs. However, in the case with multiple degrees of movement freedom, the VM is pre-allocated on multiple gNBs. Both availability of the computing resources and the quality of all potentially involved communication links are considered. To alleviate the gNBs' backhaul load for the placement of the computing resources, we restrict the list of available links exclusively to the gNBs from set $Q_t(u)$ with which the UE can communicate directly at the time t. This restriction leads to a lower overhead, as the SINR information from the gNBs in $S \backslash Q_t(u)$ is not required.

The management of VMs (starting, terminating, and adapting VMs to the actual movement and channel quality of the UEs) is done in a MEO, located in the core network [12]. Whenever the information about the UEs' movement and the channel quality is available, the MEO adapts the VM initialization based on the actual UE's velocity and position. The adaptation of the VM initialization, consists of changing the allocation of the computing resources for the UE to other gNB if the real movement of the UE differs from the predicted one. On the other gNB, the pre-allocated VM is exploited if available, otherwise the VM is started. Furthermore, the MEO terminates VMs that are no longer needed. The computing resources are allocated every $K\Delta t$ seconds to update the VM placement. The value of $K\Delta t$ can be adapted to each environment, e.g., $K\Delta t$ is set to a high value (tens of seconds) in an area with very few crossings whereas in an area with a high number of crossings, such as city center, $K\Delta t$ is set to a low value (few seconds

or less). Thus, the computing resource allocation can be dynamically adapted to various environments and various UE's mobility characteristics (walking, in a car, train, etc.).

The process of computing resource allocation is shown in Algorithm 1. The algorithm is designed for the case of UEs with several degrees of mobility freedom (line 2). At first, the UE's velocity vector is predicted (line 1). Then, if the environment is known (line 3), it is exploited to predict the closest street centers (line 4). Afterwards, the computing resources are allocated for every time instant $t + k\Delta t$ until $t + K\Delta t$ (lines 6 to 31). Only the gNBs with enough available computing resources are considered for the VM placement (lines 13 and 14).

In the next steps, the offloading metric $\alpha_\tau^w(s)$ is determined. The offloading metric $\alpha_\tau^w(s)$ is derived from the communication data rate (according to (5.12)), either from SINR map (line 17) provided that the SINR map is available (line 16), or from SINR predicted from known previous SINR levels by ARIMA (line 23) if SINR can be predicted (line 19). If SINR to the s-th gNB cannot be predicted due to a lack of information for the prediction (i.e., if SINR map is not trained or not enough known previous SINR levels are available) the offloading metric $\alpha_\tau^w(s)$ is set based on the distance $d_\tau(s, w)$ defined in (5.38) (line 20).

The sequence of the gNBs that maximizes the decision metric $\alpha_\tau^w(s)$ is selected for the VM placement $\{s_\tau^*(w)\}_t^{t+K\Delta t}$ (line 30). The sequence $\{s_\tau^*(w)\}_t^{t+K\Delta t}$ is then exploited at the MEO to manage the initialization of the VMs. The management of the VMs includes determination of the time instances when the VM is started (t_S) and ended (t_E). Between these two times, the VM on the gNBs should be up and running. The time instances t_S and t_E are derived based on $\{s_\tau^*(w)\}_t^{t+K\Delta t}$ and are equal to the first and the last occurrence of s in the sequence $\{s_\tau^*(w)\}_t^{t+K\Delta t}$, respectively. Furthermore, we avoid the pre-allocation of the VM to the gNBs, where the VM would be exploited for less than the VM startup time t_{VM} (line 33). The gNBs for which $t_E - t_S < t_{VM}$ are removed from $\{s_\tau^*(w)\}_t^{t+K\Delta t}$ and these are not considered for the VM placement. Instead, already running VMs are exploited to handle the offloading. Thus, the computing load of the gNBs is decreased and the gNBs can be exploited for the VMs of the other UEs.

An example of the Algorithm 7 determining VM pre-allocation is shown in Figure 5.19. In this example, there are three gNBs (gNB_1, gNB_2, gNB_3) and one UE located on a crossroad with three possible future directions w_1, w_2, and w_3. For each direction and each time step, the gNBs are ordered according to α (see table in the middle part in Figure 5.19). Then, t_S and t_E are determined as the first and the last occurrence of each gNB in the first row of the table in Figure 5.19 over all departure angles.

Selection of communication path

To further reduce the offloading delay, we propose also an algorithm reducing the communication delays t_O and t_C. The algorithm forces the UE to perform handover to the gNB that provides the fastest delivery of the offloaded task to the VM considering radio as well as backhaul data rates. The algorithm is inspired by our previous work, PSwH algorithm [162], [192]. The PSwH maximizes the communication data rate of the

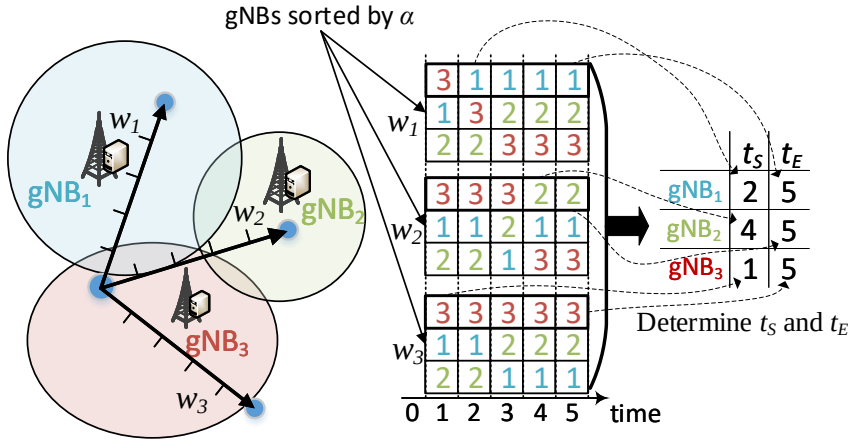


Figure 5.19. Example of VM placement by Algorithm 7 for three gNBs (number of rows in each table in the middle part of the figure) over five time instants (columns in each table in the middle part of the figure). For each departure angle, represented by individual table, a sequence (row) of the gNBs maximizing α is chosen and then exploited to determine t_S and t_E for each gNB.

UEs. However, in the PSwH, the UEs do not cooperate and the algorithm does not consider the prediction of the channel quality for resource allocation. Thus, we propose the algorithm that efficiently handles the rapid changes in the UEs' communication data rates. The selection of the communication paths for the UEs is made by an iterative update of the serving gNBs every Δt as shown in Algorithm 8, assuming fixed s_t^* for every u during given time interval $\langle t, t + \Delta t \rangle$.

The algorithm for selection of communication path starts with a determination of the serving gNBs based on the SINR of the UEs (line 1). Then, the current data rates in uplink (following (5.14)) and downlink (by adapting (5.12)) are derived from the known SINR and from the number of connected UEs. Then, a set of the gNBs $\hat{S}$ is created by sorting the gNBs in descending order based on their radio communication load $n_t^R(s)$ (line 3). The algorithm then goes through the gNBs in $\hat{S}$ that have more than one connected UE (lines 4 and 5). Four variables are defined for the communication path selection: i) minimal gain of handover to avoid exploiting handover for the UEs with a minor data rate improvement ϵ, ii) the UE u^H with the highest benefit from handover to any gNB (set initially to 0, see (line 6)), iii) the gNB s^H selected by the UE u^H as a candidate for the handover (set to 0 in initial phase), and iv) maximal achievable handover gain of all the UEs β (also set to 0 in initial phase). The algorithm iteratively searches for the UE, which can benefit the most from handover, i.e., the UE that maximizes β (lines 9 to 17). The auxiliary handover gain β^{a} (exploited to find β) is determined for each pair of the UE and the gNB that can communicate with SINR above $SINR_{min}$ (lines 9 and 10). The gain is defined as the difference between the achievable communication data rates (in both uplink and downlink) when the UE is connected to its serving gNB s_t and to a target gNB from the set $S \backslash s_t$ (line 11). If β^{a} is higher than β, u^H and s^H are updated.

Then, if β is equal to or higher than the threshold ϵ, the UE u^H is handed over to the gNB s^H (line 19) and the UEs' data rates are updated (line 20).

Complexity

The minimization of the offloading delay by the joint selection of the VM placement and the communication path leads to a combinatorial formulation. The total complexity of the DCCRA for N UEs and considering $|W|$ is $\mathcal{O}\left(N|S||Q_t|K + N|W||Q_t|K\right)$, where $|Q_t| = \max_{u \in U} |Q_\tau(u)|$. The state of the art algorithm presented in [130], further denoted as VM VM Online Approximation Placement (VM-OAP) algorithm, has complexity $\mathcal{O}\left(N|S|^2K\right)$. However, the VM-OAP algorithm is designed only for one degree of freedom. When $|W| = 1$ and all gNBs being considered for the path selection, i.e., $|Q_t| = |S|$ (worst case scenario), the DCCRA is of the same complexity as the VM-OAP algorithm in the worst case.

5.4.4 Simulation Scenario and Models

In this section, we describe simulation models and scenarios for performance evaluation carried out in MATLAB. The main simulation parameters, presented in Table 5.5, are in line with recommendations for mobile networks with small cells as defined by 3GPP in [157]. We also follow the specifications of the physical layer and frame structure parameters for LTE-A mobile networks defined in the same document. The signal propagation over radio channel is modeled according to 3GPP [] with path loss model $PL = 128.1 + 37.6log_{10}d$, where d is the distance between the UE and the gNB. We consider the mapping function between SINR and MCS defined in [176] for BER of 10 %. The minimal SINR to enable communication, $SINR_{min}$, is set to -6.9 dBm, according to [176]. We set the weighing factor for SINR map updates, χ, to 0.5 so that the SINR changes due to varying environment are quickly propagated in our model. Note that, in a real deployment, χ can be adjusted based on the environment. The backhaul of the gNBs is modeled as an optical fiber with capacities following a normal distribution with average $\mu = 100$ and variance $\sigma^2 = 2$ (in Mbit/s).

Since we target the real-time applications, the offloaded tasks with sizes of 20 and 200 kB are considered [180]. Moreover, the offloaded task with a size of 2000 kB is investigated as well, to show performance even for larger tasks. The VM startup time t_{VM}, representing the time required to initialize the VM and to prepare it to process the offloaded tasks is 4.5 s for the VMs [23]. This corresponds to the time between the moment when the VM pre-allocation begins to the moment when the offloaded application is run. Note that the t_{VM} contributes to the offloading delay only when the VM is not prepared on the gNB on time. The radio and backhaul resources are allocated to the UEs by round-robin scheduling.

The simulation area, as shown in Figure 5.20, represents a part of Prague, Czech Republic. The environment is similar to the one in [187], where the authors consider arrival and departure angles difference of 90°. In this area, four gNBs (represented by

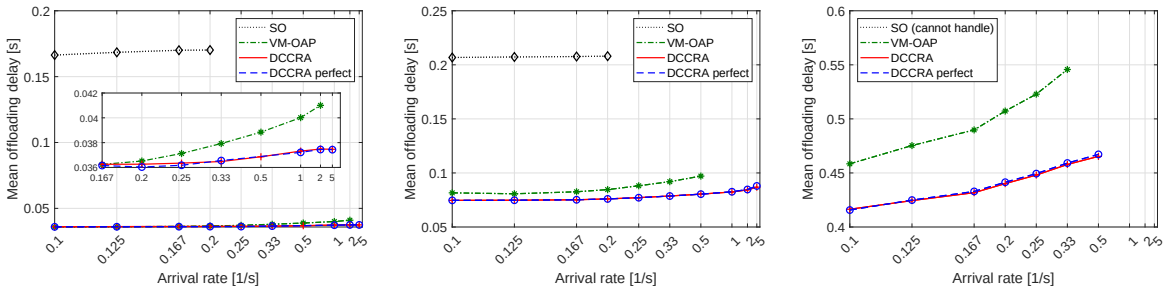
Table 5.5. Simulation parameters.

Parameter	**Value**
Simulation area	650m x 370m
Carrier frequency	2 GHz
Bandwidth of uplink/downlink	10/10 MHz
Tx power of gNB/SCgNB/UE (S_{Tx})	27/15/10 dBm
$SINR_{min}$	-6.9 dB
Weighting factor	0.5
Number of gNB/SCgNB	4/30
VM startup time t_{VM}	4.5 s
Prediction window K	200
ARIMA number of past samples	20
Offloaded task size L_{O} = results size L_{C}	20/200/2000 kB
Offloaded task number of instructions L_{P}	1e6 instructions
gNB/SCgNB CPU	3300 MIPS
Shadowing factor	6 dB
Handover interruption duration t_{HO}	30 ms
Threshold ϵ	100 kbit/s
Number of UEs	30/60/90
Speed of users	1 m/s
Backhaul capacity – Normal distribution	μ=100, σ^2=2
Simulation time/Number of simulation drops	3 600 s/ 20 drops
Simulation step	100 ms

blue discs in Figure 5.20) are deployed according to the real position of the gNBs of a mobile operator [193]. In addition, 30 SCgNBs, divided into two sets with different transmission frequencies are randomly deployed (denoted as orange crosses in Figure 5.20). To show the impact of network load on the performance, 30, 60, and 90 UEs are randomly dropped into the simulation area. The UEs follow the realistic mobility model with crossroad direction probabilities defined in [193]. The number of UEs per cell in our scenario is about 0.9, 1.8, and 2.7, which is higher than that in [130] (where they assume roughly 0.55 UEs per cell). The reason for a higher UEs' density, i.e., 60 and 90 UEs, is to evaluate the performance with highly loaded MEC servers. Furthermore, the bandwidth for communication is 10 MHz in uplink and 10 MHz in downlink allocated in FDD manner. We exploit a common handover procedure based on SINR, as described in [194], to keep the UE connected to the gNB with the highest SINR. In the simulations, the UEs move with the same speed. This might be seen as an optimistic assumption, however, the proposed DCCRA takes into account the speed of UE via (9) and (10). Thus, the DCCRA can handle easily different speeds of UEs without any degradation of performance.

The energy consumption model of the UEs follows an empirical model defined in Appendix A [149].

Figure 5.20. Simulation model with deployment of gNBs and SCgNBs.



(a) (b) (c)

Figure 5.21. Mean times required to offload, compute, and collect results of the offloaded task for 30 UEs with $L_O = L_C = 20$ kB (a), $L_O = L_C = 200$ kB (b), and $L_O = L_C = 2000$ kB (c).

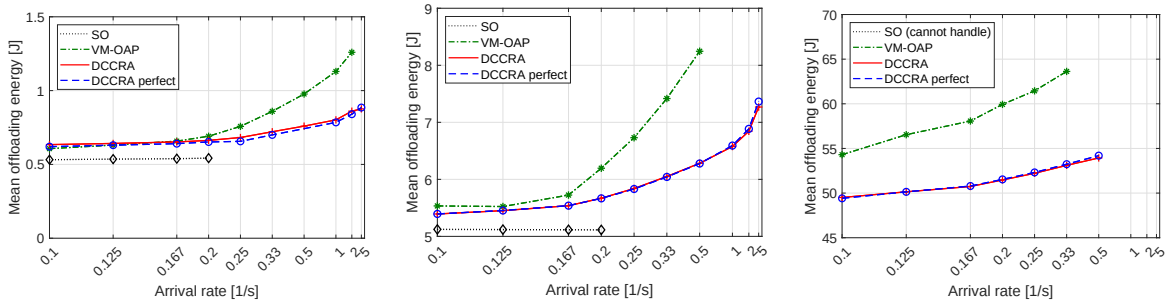
5.4.5 Performance Evaluation and Discussion of Results

The performance of the proposed algorithm (DCCRA) is compared with two state-of-the-art approaches:

- *SO* according to [175] - where the VM is kept at the serving gNB, so the VM is migrated each time handover is performed.
- *VM-OAP* according to [130] – where the VM placement is based on predicted future costs (in terms of channel quality) of its placement.

In addition to these two competitive solutions, we also show the performance of the DCCRA under perfect mobility and channel quality prediction with the VM pre-allocation on only one gNB (denoted as DCCRA-perfect in this section) to see potential improvement if the prediction would be ideal.

In Figure 5.21, we show the mean offloading delay over the task arrival rate for 30 UEs with the offloaded task size of 20 kB (Figure 5.21a), 200 kB (Figure 5.21b), and



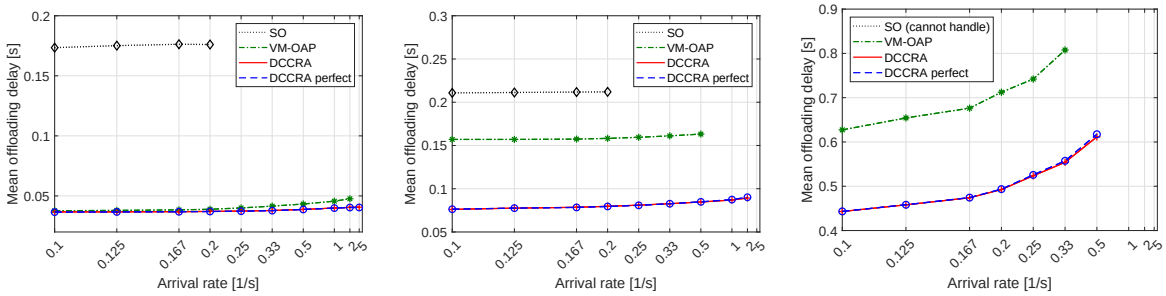
(a) (b) (c)

Figure 5.22. Mean offloading energy for 30 UEs with $L_O = L_C = 20$ kB (a), $L_O = L_C = 200$ kB (b), and $L_O = L_C = 2000$ kB (c).

2000 kB (Figure 5.21c). In each figure, we see that a higher arrival rate generally leads to a higher offloading delay because more offloaded tasks per second are generated and more communication and computation resources are consumed. Furthermore, it is shown that increasing the offloaded task size results in a higher offloading delay, as more data have to be transmitted. The SO algorithm supports offloading up to $\lambda = 0.2$ for $L_O = L_C$ up to 200 kB. The SO algorithm does not enable a higher λ as it does not exploit any prediction or pre-allocation of the VMs, thus, the resources become unavailable even for a very light computation load. The VM-OAP algorithm enables λ up to 2 for $L_O = L_C = 20$ kB. However, for a higher L_O and L_C, the VM-OAP algorithm can handle λ only up to 0.33. The VM-OAP exploits channel quality prediction, but it does not pre-allocate the VMs. The DCCRA outperforms both compared algorithms by enabling the offloading of the tasks with λ up to 5 for $L_O = L_C$ up to 200 kB, and λ up to 0.5 for $L_O = L_C = 2000$ kB. This means that the DCCRA enables offloading with almost twice higher λ than the VM-OAP. Comparing the DCCRA to the DCCRA perfect, we can see that the performance of both is very similar and the ideal prediction does not lead to any notable reduction in the offloading delay.

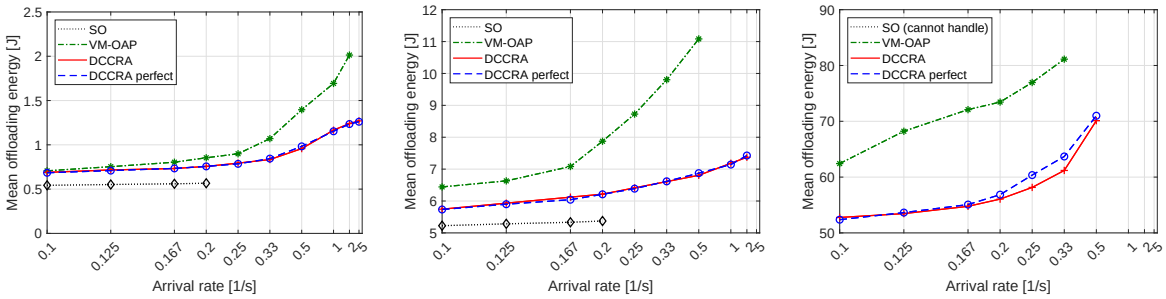
The proposed DCCRA reduces the offloading delay by up to 78 % comparing to the SO algorithm for $\lambda = 0.2$. In comparison to the VM-OAP, the DCCRA reduces the offloading delay by 8.2 % for $L_O = L_C = 20$ kB and $\lambda = 2$. Furthermore, increasing $L_O = L_C$ to 2000 kB leads to an increased gain (15.2 % for $\lambda = 0.33$) of the DCCRA in comparison to the VM-OAP algorithm. The offloading delay reduction is achieved by optimizing the placement and pre-allocation of the VMs, as well as the selection of the communication path. Increasing $L_O = L_C$ leads to a higher offloading delay for all compared algorithms, but the DCCRA keeps the offloading delay below 100 ms for small sized tasks ($L_O = L_C$ below 2000 kB).

Figure 5.22 shows the mean energy consumed by the UEs for the transmission of the offloaded task and the reception of the computing results with the offloaded task size of 20 kB (Figure 5.22a), 200 kB (Figure 5.22b), and 2000 kB (Figure 5.22c). In all these figures, any increase in λ or $L_O = L_C$, leads to an increase in the energy consumed per the offloaded task, because the network load rises. The higher energy



(a) (b) (c)

Figure 5.23. Mean times required to offload, compute, and collect results of the offloaded for 60 UEs with $L_O = L_C = 20$ kB (a), $L_O = L_C = 200$ kB (b), and $L_O = L_C = 2000$ kB (c).



(a) (b) (c)

Figure 5.24. Mean offloading energy for 60 UEs with $L_O = L_C = 20$ kB (a), $L_O = L_C = 200$ kB (b), and $L_O = L_C = 2000$ kB (c).

consumption is caused by increased communication time due to the communication and computing load of the gNBs, and the relation between the energy and the transmission time (see (A.12)). In comparison to the SO algorithm, the proposed DCCRA increases the consumed energy by less than 23 %, however, it is only 0.1 J, for $L_O = L_C = 20$ kB and 8.3 %, i.e., 0.5 J, for $L_O = L_C = 200$ kB. Note that this increase is largely compensated by a significant reduction in the offloading delay by up to 78 % and by enabling the offloading of tasks with $L_O = L_C = 2000$ kB, as shown in Figure 5.21. Furthermore, the DCCRA reduces the energy consumed for the offloading by up to 35 % compared to the VM-OAP algorithm. The reduction in the offloading energy is achieved by avoiding the overloaded communication paths and by pre-allocation of the VMs. Avoiding the overloaded communication links is done by the proposed selection of communication path, while the VMs are pre-allocated to minimize the delay of VM startup (migration). Again, we see that the DCCRA provides a similar performance as the DCCRA perfect.

In Figure 5.23, we show the mean offloading delay over the task arrival rate for 60 UEs with the offloaded task size of 20 kB (Figure 5.23a), 200 kB (Figure 5.23b), and 2000 kB (Figure 5.23c). The results follow the same trends as shown in Figure 5.21 for 30 UEs. The DCCRA increases the gain in comparison to the VM-OAP to 31 % for

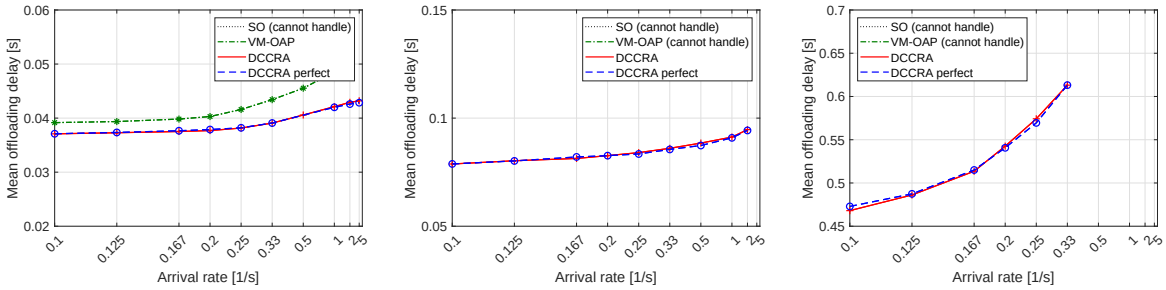
$\lambda = 0.33$ and $L_O = L_C = 2000$ kB. This is caused by the fact that the DCCRA balances the computing and communication loads among the gNBs.

Similar changes due to the increased number of UEs are seen in mean offloading energy, as shown in Figure 5.24 with the offloaded task size of a 20 kB (Figure 5.24a), 200 kB (Figure 5.24b), and 2000 kB (Figure 5.24c). However, a higher offloading delay leads to an increased energy consumption. The DCCRA leads to a similar energy consumption as the DCCRA perfect for $L_O = L_C = 200$ kB or less. In the case of $L_O = L_C = 2000$ kB, the DCCRA slightly lowers the energy consumption with respect to the DCCRA perfect. This is caused by the pre-allocation of slightly more VMs for each UE by the DCCRA comparing to the DCCRA perfect. These additional pre-allocated VMs by the DCCRA are exploited to avoid the overloaded gNBs. Note that the DCCRA perfect does not predict the number of connected UEs, thus, the predicted data rate as well as the overloading of the gNBs is not predicted perfectly. Therefore, the DCCRA provides a minor improvement over the DCCRA perfect, but at the cost of pre-allocating a higher number of VMs.

The mean offloading delay for 90 UEs is shown in Figure 5.25 with the offloaded task size of 20 kB (Figure 5.25a), 200 kB (Figure 5.25b), and 2000 kB (Figure 5.25c). The increased number of UEs, again, leads to an increased offloading delay. The SO algorithm cannot handle the offloading for 90 UEs due to keeping the VM on the serving gNB. Furthermore, the VM-OAP cannot handle the offloading for 90 UEs and $L_O = L_C$ above 20 kB, as shown in Figure 5.25b and Figure 5.25c, as it does not exploit pre-allocation. The DCCRA enables offloading with λ equal to 5, 2, and 0.5 for $L_O = L_C$ equal to 20, 200, and 2000 kB, respectively. From Figure 5.21, Figure 5.23, and Figure 5.25, we see that the DCCRA keeps the offloading delay for small offloaded tasks (below 200 kB) under 100 ms, which is not possible with any of the competitive algorithms.

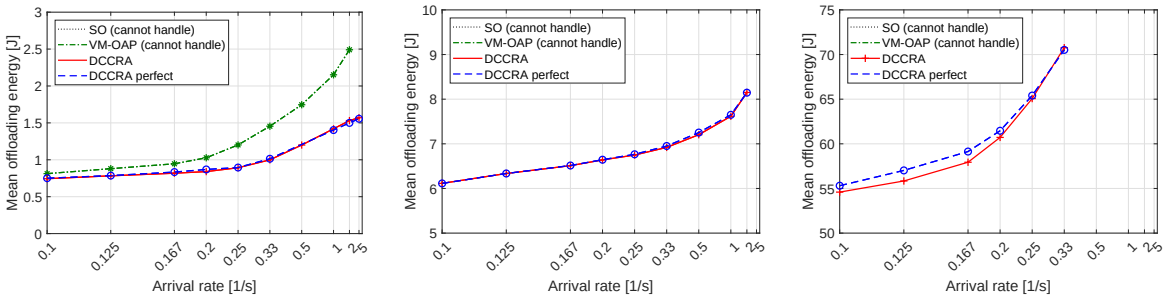
The energy consumed for the offloading for 90 UEs is shown in Figure 5.26 with the offloaded task size of 20 kB (Figure 5.26a), 200 kB (Figure 5.26b), and 2000 kB (Figure 5.26c). Again, the increased number of the UEs leads to an increased energy consumption. The results for the SO algorithm are not shown, as the algorithm cannot handle such high number of UEs. Furthermore, the VM-OAP is shown only for $L_O = L_C = 20$ kB, as it cannot handle larger offloaded task sizes with 90 UEs. The DCCRA consumes slightly less energy than the DCCRA perfect for $L_O = L_C = 2000$ kB due to the same reason as for 60 UEs (Figure 5.24c).

The mean amount of data transmitted over the backhaul due to delivery of the offloading task to the computing VM and collection of the results at the UE is shown in Figure 5.27a for 20 kB, in Figure 5.27b for 200 kB, and in Figure 5.27c for 2000 kB. Since the SO algorithm places the VMs exclusively on the serving gNB, no data is transmitted over the backhaul. Thus, the SO is not included in these figures. For the VM-OAP algorithm, the amount of data transmitted over the backhaul is constant over all investigated task arrival rates for all numbers of UEs, and for all offloaded task sizes. For the proposed DCCRA, the amount of data transmitted over the backhaul is slightly decreasing with increasing λ. This is caused by the need for a closer placement of the VMs to minimize the



(a) (b) (c)

Figure 5.25. Mean times required to offload, compute, and collect results of the offloaded for 90 UEs with $L_{\mathrm{O}} = L_{\mathrm{C}} = 20$ kB (a), $L_{\mathrm{O}} = L_{\mathrm{C}} = 200$ kB (b), and $L_{\mathrm{O}} = L_{\mathrm{C}} = 2000$ kB (c).



(a) (b) (c)

Figure 5.26. Mean offloading energy for 90 UEs with $L_{\mathrm{O}} = L_{\mathrm{C}} = 20$ kB (a), $L_{\mathrm{O}} = L_{\mathrm{C}} = 200$ kB (b), and $L_{\mathrm{O}} = L_{\mathrm{C}} = 2000$ kB (c).

communication delay when the time between two consecutive offloaded tasks is low (i.e., for a high λ). The proposed algorithm transmits 40 % less data over the backhaul comparing to the VM-OAP. This reduction is achieved by allocating the VMs in a proximity of the UEs to reduce the offloading delay and to alleviate the backhaul communication load. The DCCRA perfect transmits slightly more data over the backhaul comparing to the DCCRA. This is caused by the pre-allocation of the VM on a lower number of the gNBs, as shown in Figure 5.28. Comparing the impact of the size of offloaded task (i.e., comparing sub-figures Figure 5.27a, Figure 5.27b, and Figure 5.27c), we can see that the amount of data transmitted over the backhaul is increasing proportionally to the offloaded task size.

The number of VMs deployed for all UEs during the simulation run is shown in Figure 5.28. The sub-figures represent results for the tasks with a size of 20 kB (Figure 5.28a), 200 kB (Figure 5.28b), and 2000 kB (Figure 5.28c), respectively. To provide an insight into performance of our prior work [185], the number of pre-allocated VMs is equal to the number of UEs (60) multiplied by the number of gNBs (34), i.e., 2040 VMs in our scenario. This number is many times higher than the number of gNBs where the VM is pre-allocated by the DCCRA, thus, we do not show the lines for 2040 VMs in the figure. Since only our proposed algorithm exploits the possibility to deploy more than one VM

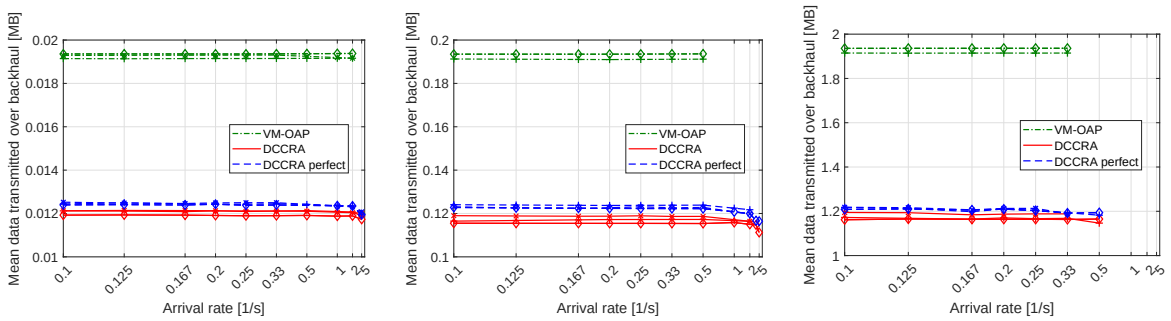


Figure 5.27. Mean amount of data transmitted per backhaul per task for $L_O = L_C = 20$ kB (a), $L_O = L_C = 200$ kB (b), and $L_O = L_C = 2000$ kB (c), solid line represents 30 UEs, dashed 60 UEs and dotted 90 UEs.

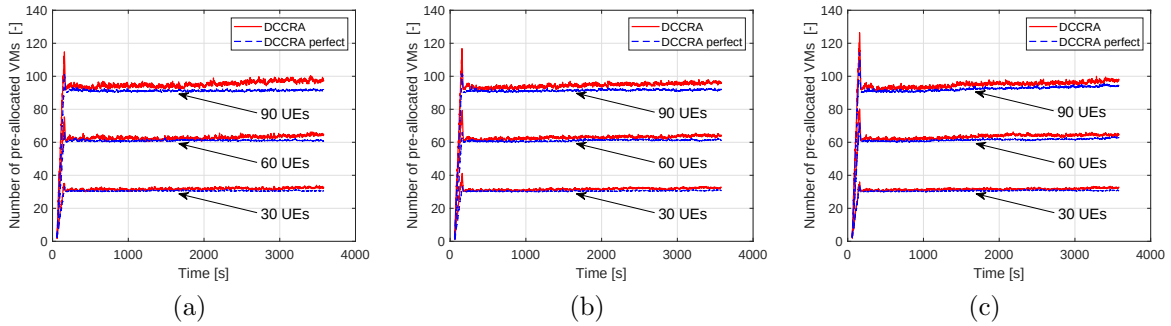


Figure 5.28. Number of pre-allocated VMs for the proposal during the simulation run. for $L_O = L_C = 20$ kB (a), $L_O = L_C = 200$ kB (b), and L_O=$L_C = 2000$ kB (c), solid line represents 30 UEs, dashed 60 UEs and dotted 90 UEs.

per UE, other algorithms are not depicted. The SO and VM-OAP algorithms deploy the same number of VMs as the number of UEs offloading their tasks, i.e., 30 VMs for 30 UEs, 60 VMs for 60 UEs, and 90 VMs for 90 UEs. To show the impact of mobility and channel prediction, we compare the DCCRA to the DCCRA perfect for $\lambda = 2$ with 30, 60, and 90 UEs. From the Figure 5.28, we see that just after the simulation starts, there is a steep increase in the number of the deployed VMs, as the number of UEs offloading their tasks increases. However, when all the UEs are offloading their tasks, the number of deployed VMs stabilizes at 32.4 VMs for 30 UEs, 65 VMs for 60 UEs, and 98 VMs for 90 UEs. In case of the DCCRA perfect, the number of deployed VMs is 31 VMs for 30 UEs, 61.5 VMs for 60 UEs, and 93 VMs for 90 UEs. On the average, there are 1.08 and 1.03 VMs pre-allocated per UE for the DCCRA and the DCCRA perfect, respectively, for all sizes of the offloaded tasks. The difference in the number of pre-allocated VMs between the DCCRA and the DCCRA perfect is caused by the need to pre-allocate more VMs for the DCCRA to compensate for the mobility and channel prediction inaccuracies. The minor fluctuation in the number of pre-allocated VMs over time is caused by the fact that the UEs are selecting from multiple future angles at irregular time instants.

With pre-allocation of 8 % more VMs than the number of UEs, the DCCRA enables the offloading of the real-time task with very high arrival rate.

5.4.6 Conclusion

In this section, we have proposed a novel algorithm for dynamic pre-allocation of computing and communication resources for the MEC. The algorithm dynamically pre-allocates VMs considering the computation load of gNBs and selects the best communication path between the UE and the gNB with allocated VM. For the proposed algorithm, we have designed a suitable mobility channel prediction with a low complexity.

Comparing to state of the art approaches, the proposed algorithm reduces the offloading delay by up to 64 %, while reducing UE's energy consumption by up to 39 %. The proposed algorithm enables offloading of tasks with arrival rate up to 5 tasks per second per UE for small task sizes. The competitive algorithms do not surpass 2 and 0.5 tasks per second for very small and small task sizes. The proposed algorithm, also provides offloading delay below 100 ms for small sized offloaded tasks, making it suitable for real-time offloading. Furthermore, we show that the performance of the proposed algorithm is similar to the case with perfect mobility and channel prediction.

Algorithm 7 Allocation of computing resources.

Calculate UE's Δx and Δy velocity via (5.24) and (5.25)
for $w \in W$ such that $P(w|v,(x_t,y_t)) > 0$ **do**
  if there exists a known map of street centers **then**
    Calculate j_w^* via (5.35) from (5.33) and (5.34)
  end if
  for $\tau = t, t+\Delta t, \ldots, t+K\Delta t$ **do**
    if there exists a known map of street centers **then**
      Calculate (x_τ^w, y_τ^w) via (5.36) and (5.37)
    else
      Calculate (x_τ^w, y_τ^w) via (5.33) and (5.34)
    end if
    for $s \in Q_\tau(u)$ **do**
      if $\omega_\tau(s) < \omega(u)$ **then**
        $Q_\tau(u) \leftarrow Q_\tau(u) \backslash s$
      else
        if $\Psi_{x,y} \neq 0, \forall \lfloor x_\tau^w \rceil, \lfloor y_\tau^w \rceil$ **then**
          $SINR_\tau(s) \leftarrow \Psi_{x_\tau^w, y_\tau^w}$
        else
          if $SINR_\tau(s)$ is not predictable **then**
            $\alpha_\tau^w(s) \leftarrow \frac{1}{d_\tau(s,w)}, \forall s \in Q_\tau(u)$
            break
          else
            Predict SINR by (5.43)
          end if
        end if
        Calculate $c_\tau(u,s)$ via (5.12)
        $\alpha_\tau^w(s) \leftarrow c_\tau(u,s)$
      end if
    end for
    $s_\tau^*(w) \leftarrow \arg\max_{s \in Q_\tau(u)} \alpha_\tau^w(s)$
  end for
end for
Remove s with VM exploited for less than t_{VM}

Algorithm 8 Allocation of communication resources.

Determine serving gNBs maximizing SINR.
Calculate the uplink and downlink data rates $c_t^{UL}(u, s, s^*)$ and $c_t^{DL}(u, s, s^*)$ for each UE and gNB.
Sort the gNBs in descending order based on $n_t^R(s)$ to $\hat{S}$
for do $\hat{s} \in \hat{S}$
  if then $n_t^R(\hat{s}) > 1$
    Initialize $u^H = 0$, and $s^H = 0$.
    while do(true)
      $\beta = 0$
      for do $u \in U$ such that $s_t = \hat{s}$
        for do $s \in Q_\tau(u)$
          $\beta^{\mathrm{a}} = min\big(c_t^{UL}(u, s, s^*) - c_t^{UL}(u, \hat{s}, s^*),$
            $c_t^{DL}(u, s, s^*) - c_t^{DL}(u, \hat{s}, s^*)\big)$
          if then $\beta^{\mathrm{a}} > \beta$
            $u^H = u$, $s^H = s$.
            $\beta = \beta^{\mathrm{a}}$
          end if
        end for
      end for
      if then $\beta \geq \epsilon$
        $s_t(u^H) = s^H$
        Update $c_t^{UL}(u, s, s^*)$ and $c_t^{DL}(u, s, s^*)$.
      else
        break
      end if
    end while
  end if
end for

Chapter 6

Conclusion

This thesis focuses on collection of mobile network information from the UEs and allocation of communication and computation resources for the MEC services. In this chapter, a summary of the thesis is provided, followed by a description of the research contribution based on the presented work. To conclude this chapter, a future research direction is provided.

6.1 Thesis summary

This thesis focuses on allocation of communication and computation resources for MEC. The motivation behind the thesis is to design and propose solution for real-time self optimization of the mobile networks. To achieve this goal, several objectives are specified, and solutions are provided.

The first part of this thesis describes solutions for collecting a large amount of data from mobile devices (UEs, sensors, vehicles, etc.). One of the main limitations of the mobile networks is number of devices that can be served by the mobile networks. Thus, we have proposed a solution to overcome current limitation to enable more than 65 000 devices per base station transmitting small amounts of data (tens of bits), which is in line with the expected number of devices connected to one gNB in 5G (10 000 to 100 000 devices, see [2]). The solution for enabling such a huge number of devices exploits D2D for data relaying. However, it is necessary to provide a solution for allocation of the communication resources to satisfy the mobile users. Thus, we solve this via NBS that leads to a natural cooperation of the devices. Moreover, the proposed solution is in the closed form, therefore, have a very low computation complexity and works even under fast changing communication quality. An important aspect for the devices exploiting the mobile networks is the energy consumption. Therefore, we provide an analysis of energy consumption for data relaying via D2D communication. The analysis shows, that the D2D relaying lowers energy consumption of the devices in comparison to the traditional communication, where all transmissions go directly to the BS.

The collected mobile network information is necessary for deployment of the FlyBSs that provide an option to tackle the problem of the time and space varying requirements

on the mobile networks. This is achieved by positioning the FlyBSs to satisfy the requirements in the mobile networks and repositioning when the time and space requirements change. To this end, we propose a solution that jointly positions the FlyBSs and associates the UEs. The proposed solution significantly (up to 30%) increases the UEs satisfaction with the provided data rates.

The second part of the thesis focuses on allocation of the communication and computation resources to the mobile users exploiting the MEC services (including distributed computing). First, we propose a solution, based on the MDP, for allocation of the communication resources, that selects the communication path, i.e., serving BS. The proposed solution exploits the BSs in proximity of the user and selects the best communication path based on the user's preferences for delay and energy consumption. It is shown, that this approach can significantly reduce communication delay or energy compared to the state of the art solutions. Compared to the existing approach, the proposed solution reduces communication delay by up to 29%.

The allocation of the communication resources is extended to jointly allocate communication and computation resources by the proposed algorithm. First, we propose a solution that considers UEs mobility with a known prediction accuracy. The proposed solution reduces offloading delay in comparison to the state of the art solutions by 10-66%. This solution is further extended so the prediction accuracy is unknown. The proposed solution exploits existing mobility prediction algorithms to select which computation resources should be used for processing of the offloaded task and to select the most suitable communication path. The proposal is formed from two cooperating algorithms in order to achieve a low complexity and to enable offloading of tasks with a high frequency, which is not possible with the existing approaches. The proposed solution reduces offloading delay by up to 66% compared to the state of the art algorithms and enable offloading of offloaded tasks with time between two tasks of 1s.

6.2 Research contributions

The solutions proposed in the thesis have been presented at multiple conference and journal papers indexed in WoS. The research contributions of these papers and related objectives in each paper are as follows:

- A solution to enable collection of data from a large number of devices had been proposed. The proposed solution combines reduction of the communication overhead and exploitation of buffering (transmission of multiple data from a single device) and clustering (exploiting D2D for relayed transmissions). This solution, described in Section 4.1 and published in [132], fulfills the Objective 1.

- In Section 4.2, we propose time resource allocation algorithm based on NBS so that the UEs are motivated to cooperate in relaying communication. This is necessary for enabling communication of a huge number of UEs. This solution, described in Section 4.2 and published in [133], fulfills the Objective 2.

- An analysis of energy consumed for the D2D relaying and benefits of the relaying, that is exploited for communication of a huge number of UEs are presented in Section 4.3. This solution, described in Section 4.3 and published in [134], fulfills the Objective 2.

- Algorithms for joint positioning of the FlyBSs and association of the UEs to improve their satisfaction with the provided QoS are proposed. These algorithms are based on the GA and the PSO and exploit the collected mobile network information from the UEs. This solution, described in Section 4.4 and published in [135], fulfills the Objective 3.

- An algorithm for selection of communication path between a UE and MEC hosts has been proposed. The proposed algorithm, based on the MDP exploits possibility of handover to shorten the time of transfer of data for computation by avoiding usage of low capacity backhaul. This solution, described in Section 5.2 and published in [162, 192], fulfills the Objective 4.

- A solution for joint communication and computation resource allocation based on predicted mobility of users and load of eNBs' communication and computation resources, that consists of two cooperative algorithms has been proposed. The first proposed algorithm, dynamic VM placement, decides whether there is a more suitable place for the VM allocation before the offloaded task is processed at the VM. The second algorithm is path selection enhanced by mobility prediction. This solution, described in Section 5.3 and published in [185], fulfills the Objective 5.

- We further extended previous work by proposing a low-complexity computing and communication resource allocation for offloading of real-time computing tasks generated with a high arrival rate by the mobile users. In comparison to previous work, we assume that the mobility prediction accuracy is unknown and the VMs are not prepared on all eNBs to reduce the computation load. This solution, described in Section 5.4 and published in [195], fulfills the Objective 5.

6.3 Future research direction

This thesis focuses on allocation of communication and computation resources for the MEC. The thesis consists of two parts, collection of the mobile network information and allocation of communication and computation resources in the MEC.

In the collection of the mobile network information, primary focus is given to static users. Therefore, this work should be extended to consider mobility of the users and the consecutive challenges, such as handovers or fast varying channel quality. To overcome this, the algorithms for selection and management of the clusters (one device relaying data from others with D2D) should be adapted. To this end, it is also necessary to extend the time resource allocation based on the NBS in the frequency domain to accommodate

even higher number of UEs. The exploitation of the FlyBSs that provides improves UEs satisfaction with the provided data rates should be extended to consider the mobility of the UEs as well. Moreover, due to the limited flying time of the FlyBS the proposed solution should be extended to consider energy consumption of the FlyBSs.

The allocation of the communication and computation resources based on the collected mobile network information provides solutions for mobile UEs even without knowledge of the future movement. The work presented in [195] enables real-time computation offloading. However, with novel applications and services, the offloading delay should be reduced even further. Therefore, mmWaves or Visible Light Communication should be considered for lowering communication delay. Due to costly deployment of the ultra-dense cells, exploitation of the relayed communication for the MEC can provide a feasible option in the areas with a poor coverage. Apart from reducing delay, the energy consumption of communication is critical aspect. Thus, the proposed algorithms should not only consider energy consumption, but target on reduction of the energy consumption of the UEs to increase the battery life time when the MEC services are being used.

With mobile networks shifting from the specialized hardware towards software defined networks, the mobile networks become more flexible and the proposed solution can exploit technologies such as Open Air Interface (OAI) [196] or Open Radio Access Network (O-RAN) [197], which run mobile networks on generic hardware. Therefore, the proposes solution should be implemented and tested in the real mobile networks to validate their performance.

References

[1] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. Soong, and J. C. Zhang, "What will 5G be?" *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1065–1082, 2014.

[2] NetWorld2020, "5G: Challenges,Research Priorities, and Recommendations," *Joint White Paper*, September 2014.

[3] A. Imran, A. Zoha, and A. Abu-Dayya, "Challenges in 5G: How to Empower SON with Big Data for Enabling 5G," *IEEE Network*, vol. 28, no. 6, pp. 27–33, 2014.

[4] 3GPP, "Telecommunication management; Study on the Self-Organizing Networks (SON) for 5G networks," 3rd Generation Partnership Project (3GPP), TR, December 2019.

[5] L. Cao, J. Zhang, and N. Kanno, "Multi-user Cooperative Communications with Relay-coding for Uplink IMT-advanced 4G Systems," in *IEEE Global Telecommunications Conference (GLOBECOM)*. IEEE, 2009, pp. 1–6.

[6] M. N. Tehrani, M. Uysal, and H. Yanikomeroglu, "Device-to-device Communication in 5G Cellular Networks: Challenges, Solutions, and Future Directions," *IEEE Communications Magazine*, vol. 52, no. 5, pp. 86–92, 2014.

[7] Z. Han, Z. Ji, and K. R. Liu, "Fair Multiuser Channel Allocation for OFDMA Networks Using Nash Bargaining Solutions and Coalitions," *IEEE Transactions on Communications*, vol. 53, no. 8, 2005.

[8] A. Leshem and E. Zehavi, "Game Theory and the Frequency Selective Interference Channel," *IEEE Signal Processing Magazine*, vol. 26, no. 5, 2009.

[9] L. Xu, A. Nallanathan, J. Yang, and W. Liao, "Power and Bandwidth Allocation for Cognitive Heterogeneous Multi-homing Networks," *IEEE Transactions on Communications*, 2017.

[10] Z. Becvar, M. Vondra, P. Mach, J. Plachy, and D. Gesbert, "Performance of Mobile Networks with UAVs: Can Flying Base Stations Substitute Ultra-dense Small Cells?" in *European Wireless*, 2017, pp. 1–7.

[11] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Optimal Transport Theory for Cell Association in UAV-enabled Cellular Networks," *IEEE Communications Letters*, vol. 21, no. 9, pp. 2053–2056, 2017.

[12] European Telecommunications Standards Institute (ETSI), "GS MEC 003 V1.1.1, Version 3.2.0 Mobile Edge Computing (MEC); Framework and Reference Architecture," European Telecommunications Standards Institute (ETSI), Tech. Rep., 2016.

[13] Y. Wang, W. Shi, and M. Hu, "Virtual Servers Co-Migration for Mobile Accesses: Online versus Off-Line," *IEEE Transactions on Mobile Computing*, vol. 14, no. 12, pp. 2576–2589, 2015.

[14] B. I. Ismail, E. Mostajeran Goortani, M. B. Ab Karim, W. Ming Tat, S. Setapa, J. Y. Luke, and O. Hong Hoe, "Evaluation of Docker as Edge Computing Platform," in *IEEE Conference on Open Systems (ICOS)*. IEEE, 2015, pp. 130–135.

[15] S. Abolfazli, Z. Sanaei, E. Ahmed, A. Gani, and R. Buyya, "Cloud-based Augmentation for Mobile Devices: Motivation, Taxonomies, and Open Challenges," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 1, pp. 337–368, 2014.

[16] F. Giust, G. Verin, K. Antevski, J. Chou, Y. Fang, W. Featherstone, F. Fontes, D. Frydman, A. Li, A. Manzalini *et al.*, "MEC Deployments in 4G and Evolution Towards 5G," *ETSI White Paper*, vol. 24, pp. 1–24, 2018.

[17] M. V. Barbera, S. Kosta, A. Mei, and J. Stefa, "To Offload or not to Offload? the Bandwidth and Energy Costs of Mobile Cloud Computing," in *IEEE International Conference on Computer Communications (INFOCOM)*. IEEE, 2013, pp. 1285–1293.

[18] B. Addis, D. Ardagna, A. Capone, and G. Carello, "Energy-aware Joint Management of Networks and Cloud Infrastructures," *Computer Networks*, vol. 70, pp. 75–95, 2014.

[19] F. Lobillo, Z. Becvar, M. A. Puente, P. Mach, F. L. Presti, F. Gambetti, M. Goldhamer, J. Vidal, A. K. Widiawan, and E. C. Strinati, "An Architecture for Mobile Computation Offloading on Cloud-enabled LTE Small Cells," in *IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*. IEEE, 2014, pp. 1–6.

[20] P. Mach and Z. Becvar, "Mobile Edge Computing: A Survey on Architecture and Computation Offloading," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 3, pp. 1628–1656, 2017.

[21] V. Di Valerio and F. L. Presti, "Optimal Virtual Machines Allocation in Mobile Femto-cloud Computing: An MDP Approach," in *IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*. IEEE, 2014, pp. 7–11.

[22] H. Maziku and S. Shetty, "Network Aware VM Migration in Cloud Data Centers," in *IEEE GENIResearch and Educational Experiment Workshop (GREE)*. IEEE, 2014, pp. 25–28.

[23] K. Razavi, G. V. D. Kolk, and T. Kielmann, "Prebaked μVMs: Scalable, Instant VM Startup for IaaS Clouds," in *IEEE International Conference on Distributed Computing Systems*, vol. 2015-July. IEEE, 2015, pp. 245–255.

[24] H. Holma and A. Toskala, *LTE for UMTS-OFDMA and SC-FDMA based radio access*. John Wiley & Sons, 2009.

[25] R.-G. Cheng, C.-H. Wei, S.-L. Tsao, and F.-C. Ren, "RACH Collision Probability for Machine-type Communications," in *IEEE Vehicular Technology Conference (VTC Spring)*. IEEE, 2012, pp. 1–5.

[26] Nokia, "What is Going on in Mobile Broadband Networks? Smartphone Traffic Analysis and Solutions," *Nokia Networks white paper*, 2015.

[27] SK Telecom, "SK Telecom's View on 5G Vision, Architecture, Technology, and Spectrum," *5G White Paper*, 2014.

[28] Y. Zeng, R. Zhang, and T. J. Lim, "Wireless Communications with Unmanned Aerial Vehicles: Opportunities and Challenges," *IEEE Communications Magazine*, vol. 54, no. 5, pp. 36–42, 2016.

[29] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2, (Release 12)," 3rd Generation Partnership Project (3GPP), TS, 2015.

[30] D. Dimitrova, H. Van Den Berg, R. Litjens, and G. Heijenk, "Scheduling Strategies for LTE Uplink with Flow Behaviour Analysis," in *ERCIM Workshop on eMobility*. Citeseer, 2010, pp. 15–26.

[31] 3GPP, "Technical Specification Group Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA);Physical layer procedures," 3rd Generation Partnership Project (3GPP), TS, 2016.

[32] W. Nitzold, D. Drajic, S. Saur, N. Ognjanovic, C. Abgrall, E. C. Strinati, D. Ktenas, P. Bhat, B. Devillers, G. Cocco *et al.*, "Final Report on LTE-M Algorithms and Procedures," *Large Scale Integrating Project, EXALTED-Expanding LTE for Devices, Tech. Rep. Deliverable ID: WP3 D*, vol. 3, 2012.

[33] R. Ratasuk, N. Mangalvedhe, A. Ghosh, and B. Vejlgaard, "Narrowband LTE-M System for M2M Communication," in *IEEE Vehicular Technology Conference (VTC2014-Fall)*. IEEE, 2014, pp. 1–5.

[34] N. Michailow, M. Matthé, I. S. Gaspar, A. N. Caldevilla, L. L. Mendes, A. Festag, and G. Fettweis, "Generalized Frequency Division Multiplexing for 5th Generation Cellular Networks," *IEEE Transactions on Communications*, vol. 62, no. 9, pp. 3045–3061, 2014.

[35] R.-G. Cheng, J. Chen, D.-W. Chen, and C.-H. Wei, "Modeling and Analysis of an Extended Access Barring Algorithm for Machine-type Communications in LTE-A Networks," *IEEE Transactions on Wireless Communications*, vol. 14, no. 6, pp. 2956–2968, 2015.

[36] K. Zhou and N. Nikaein, "Packet Aggregation for Machine Type Communications in LTE with Random Access Channel," in *IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2013, pp. 262–267.

[37] C. Ubeda, S. Pedraza, M. Regueira, and J. Romero, "LTE FDD Physical Random Access Channel Dimensioning and Planning," in *IEEE Vehicular Technology Conference (VTC Fall)*. IEEE, 2012, pp. 1–5.

[38] N. Abu-Ali, A.-E. M. Taha, M. Salah, and H. Hassanein, "Uplink Scheduling in LTE and LTE-advanced: Tutorial, Survey and Evaluation Framework," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 3, pp. 1239–1265, 2013.

[39] S. Ye, S. H. Wong, and C. Worrall, "Enhanced Physical Downlink Control Channel in LTE Advanced Release 11," *IEEE Communications Magazine*, vol. 51, no. 2, pp. 82–89, 2013.

[40] J. Puttonen, T. Henttonen, N. Kolehmainen, K. Aschan, M. Moisio, and P. Kela, "Voice-over-IP Performance in UTRA Long Term Evolution Downlink," in *IEEE Vehicular Technology Conference (VTC Spring)*. IEEE, 2008, pp. 2502–2506.

[41] K. Pradap, V. Ramachandran, and S. Kalyanasundaram, "Uplink Buffer Status Reporting for Delay Constrained Flows in 3GPP Long Term Evolution," in *IEEE Wireless Communications and Networking Conference*. IEEE, 2009, pp. 1–6.

[42] 3GPP, "Medium Access Control (MAC) protocol specification," 3rd Generation Partnership Project (3GPP), TS, 2015. [Online]. Available: http://www.3gpp.org/dynareport/36321.htm

[43] R. Fracchia, C. Gomez, and A. Tripodi, "R-RoHC: a Single Adaptive Solution for Header Compression," in *IEEE Vehicular Technology Conference (VTC Spring)*. IEEE, 2011, pp. 1–5.

[44] J. Yunjie, L. Ming, Z. Song, and D. Pengtao, "A Clustering Routing Algorithm Based on Energy and Distance in WSN," in *International Conference on Computer Distributed Control and Intelligent Environmental Monitoring (CDCIEM)*. IEEE, 2012, pp. 9–12.

[45] Q. Wu, G. Li, W. Chen, and D. W. K. Ng, "Energy-Efficient D2D Overlaying Communications with Spectrum-Power Trading," *IEEE Transactions on Wireless Communications*, 2017.

[46] L. Xu, C. Jiang, Y. Shen, T. Q. Quek, Z. Han, and Y. Ren, "Energy Efficient D2D Communications: A Perspective of Mechanism Design," *IEEE Transactions on Wireless Communications*, vol. 15, no. 11, 2016.

[47] A. Bletsas, A. Khisti, D. P. Reed, and A. Lippman, "A Simple Cooperative Diversity Method Based on Network Path Selection," *IEEE Journal on selected areas in communications*, vol. 24, no. 3, 2006.

[48] I. Krikidis, J. Thompson, S. McLaughlin, and N. Goertz, "Amplify-and-forward with Partial Relay Selection," *IEEE Communications Letters*, vol. 12, no. 4, pp. 235–237, April 2008.

[49] Y. Jing and H. Jafarkhani, "Network Beamforming Using Relays With Perfect Channel Information," *IEEE Transactions on Information Theory*, vol. 55, no. 6, pp. 2499–2517, June 2009.

[50] S. Ikki and M. Ahmed, "Performance Analysis of Adaptive Decode-and-forward Cooperative Diversity Networks with Best-relay Selection," *IEEE Transactions on Communications*, vol. 58, no. 1, pp. 68–72, January 2010.

[51] A. Kalantari, M. Mohammadi, and M. Ardebilipour, "Performance Analysis of Opportunistic Relaying over Imperfect Non-identical Log-normal Fading Channels," in *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, Sept 2011, pp. 1909–1913.

[52] A. Nosratinia, T. Hunter, and A. Hedayat, "Cooperative Communication in Wireless Networks," *IEEE Communications Magazine*, vol. 42, no. 10, pp. 74–80, Oct 2004.

[53] Q. Li, R. Hu, Y. Qian, and G. Wu, "Cooperative Communications for Wireless Networks: Techniques and Applications in LTE-advanced Systems," *IEEE Wireless Communications*, vol. 19, no. 2, April 2012.

[54] M. Naeem, A. Anpalagan, M. Jaseemuddin, and D. C. Lee, "Resource Allocation Techniques in Cooperative Cognitive Radio Networks," *IEEE communications surveys & tutorials*, vol. 16, no. 2, 2014.

[55] A. Asadi, V. Mancuso, and R. Gupta, "DORE: An Experimental Framework to Enable Outband D2D Relay in Cellular Networks," *IEEE/ACM Transactions on Networking*, vol. 25, no. 5, pp. 2930–2943, Oct 2017.

[56] A. Al-Hourani, S. Kandeepan, and E. Hossain, "Relay-Assisted Device-to-Device Communication: A Stochastic Analysis of Energy Saving," *IEEE Transactions on Mobile Computing*, vol. 15, no. 12, pp. 3129–3141, Dec 2016.

[57] A. Asadi and V. Mancuso, "Network-Assisted Outband D2D-Clustering in 5G Cellular Networks: Theory and Practice," *IEEE Transactions on Mobile Computing*, vol. 16, no. 8, pp. 2246–2259, Aug 2017.

[58] Y. Li, J. Li, J. Jiang, and M. Peng, "Performance Analysis of Device-to-device Underlay Communication in Rician Fading Channels," in *IEEE Global Communications Conference (Globecom 2013)*, Dec 2013, pp. 4465–4470.

[59] D. D. Penda, N. Nomikos, T. Charalambous, and M. Johansson, "Minimum Power Scheduling under Rician Fading in Full-Duplex Relay-Assisted D2D Communication," in *IEEE Globecom Workshops (GC Wkshps)*, Dec 2017, pp. 1–6.

[60] D. D. Penda, R. S. Risuleo, P. E. Valenzuela, and M. Johansson, "Optimal Power Control for D2D Communications under Rician Fading: A Risk Theoretical Approach," in *IEEE Global Communications Conference (Globecom)*, Dec 2017, pp. 1–6.

[61] C. Kai, H. Li, L. Xu, Y. Li, and T. Jiang, "Energy-Efficient Device-to-Device Communications for Green Smart Cities," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 4, pp. 1542–1551, April 2018.

[62] M. Usman, M. R. Asghar, I. S. Ansari, M. Qaraqe, and F. Granelli, "An Energy Consumption Model for WiFi Direct Based D2D Communications," in *IEEE Global Communications Conference (Globecom)*, 2018, pp. 1–6.

[63] G. G. Messier, "Opportunistic Transmission using Large Scale Channel Effects," *IEEE Transactions on Communications*, vol. 58, no. 11, pp. 3110–3114, November 2010.

[64] L. Song, D. Niyato, Z. Han, and E. Hossain, "Game-theoretic Resource Allocation Methods for Device-to-Device Communication," *IEEE Wireless Communications*, vol. 21, no. 3, pp. 136–144, 2014.

[65] F. Wang, C. Xu, L. Song, Q. Zhao, X. Wang, and Z. Han, "Energy-Aware Resource Allocation for Device-to-Device Underlay Communication," *IEEE international conference on communications (ICC)*, 2013.

[66] M. Elhattab, M. M. Elmesalawy, and I. I. Ibrahim, "A Game Theoretic Framework for Device Association in Heterogeneous Cellular Networks With H2H/IoT Co-Existence," *IEEE Communications Letters*, vol. 21, no. 2, 2017.

[67] T. Liu and G. Wang, "Resource Allocation for Device-to-device Communications as an Underlay Using Nash Bargaining Game Theory," *International Conference on Information and Communication Technology Convergence (ICTC)*, 2015.

[68] B. Duan, Y. Cai, J.-C. Zheng, and W. Yang, "Energy-efficient Resource Allocation in Cooperative Wireless Networks using Nash Bargaining Solution," 2013.

[69] M. Kamel, W. Hamouda, and A. Youssef, "Ultra-dense Networks: A Survey," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 4, 2016, best paper award.

[70] X. Lin, V. Yajnanarayana, S. D. Muruganathan, S. Gao, H. Asplund, H.-L. Maattanen, M. Bergstrom, S. Euler, and Y.-P. E. Wang, "The Sky is not the Limit: LTE for Unmanned Aerial Vehicles," *IEEE Communications Magazine*, vol. 56, no. 4, pp. 204–210, 2018.

[71] J. Chen, Q. Wu, Y. Xu, Y. Zhang, and Y. Yang, "Distributed Demand-Aware Channel-Slot Selection for Multi-UAV Networks: A Game-Theoretic Learning Approach," *IEEE Access*, vol. 6, pp. 14 799–14 811, 2018.

[72] M. Mozaffari, W. Saad, M. Bennis, Y.-H. Nam, and M. Debbah, "A Tutorial on UAVs for Wireless Networks: Applications, Challenges, and Open Problems," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 3, pp. 2334–2360, 2019.

[73] J. Chen and D. Gesbert, "Optimal Positioning of Flying Relays for Wireless Networks: A LOS Map Approach," in *IEEE International Conference on Communications (ICC)*, 2017.

[74] M. M. Azari, F. Rosas, K.-C. Chen, and S. Pollin, "Ultra Reliable UAV Communication Using Altitude and Cooperation Diversity," *IEEE Transactions on Communications*, vol. 66, no. 1, pp. 330–344, 2018.

[75] M. Alzenad, A. El-Keyi, and H. Yanikomeroglu, "3D Placement of an Unmanned Aerial Vehicle Base Station for Maximum Coverage of Users with Different QoS Requirements," *IEEE Wireless Communications Letters*, 2017.

[76] R. Fan, J. Cui, S. Jin, K. Yang, and J. An, "Optimal Node Placement and Resource Allocation for UAV Relaying Network," *IEEE Communications Letters*, vol. 22, no. 4, pp. 808–811, 2018.

[77] Y. J. A. Zhang, L. Qian, J. Huang *et al.*, "Monotonic Optimization in Communication and Networking Systems," *Foundations and Trends® in Networking*, vol. 7, no. 1, pp. 1–75, 2013.

[78] P. Yang, X. Cao, C. Yin, Z. Xiao, X. Xi, and D. Wu, "Proactive Drone-cell Deployment: Overload Relief for a Cellular Network Under Flash Crowd Traffic," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 10, pp. 2877–2892, 2017.

[79] Y. Cao, N. Zhao, F. R. Yu, M. Jin, Y. Chen, J. Tang, and V. C. M. Leung, "Optimization or Alignment: Secure Primary Transmission Assisted by Secondary Networks," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 4, pp. 905–917, April 2018.

[80] C.-F. Huang and Y.-C. Tseng, "The Coverage Problem in a Wireless Sensor Network," *Mobile networks and Applications*, vol. 10, no. 4, 2005.

[81] H. Huang and A. V. Savkin, "An Algorithm of Efficient Proactive Placement of Autonomous Drones for Maximum Coverage in Cellular Networks," *IEEE Wireless Communications Letters*, 2018.

[82] X. Zhang and L. Duan, "Fast Deployment of UAV Networks for Optimal Wireless Coverage," *IEEE Transactions on Mobile Computing*, 2018.

[83] X. Wang, G. Xing, Y. Zhang, C. Lu, R. Pless, and C. Gill, "Integrated Coverage and Connectivity Configuration in Wireless Sensor Networks," in *ACM Int. Conference on Embedded Networked Sensor Systems*, 2003.

[84] H. Ghazzai, E. Yaacoub, M.-S. Alouini, Z. Dawy, and A. Abu-Dayya, "Optimized LTE Cell Planning with Varying Spatial and Temporal User Densities," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 3, 2016.

[85] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Efficient Deployment of Multiple Unmanned Aerial Vehicles for Optimal Wireless Coverage," *IEEE Communications Letters*, vol. 20, no. 8, 2016.

[86] F. Lagum, I. Bor-Yaliniz, and H. Yanikomeroglu, "Strategic Densification with UAV-BSs in Cellular Networks," *IEEE Wireless Communications Letters*, vol. 7, no. 3, pp. 384–387, 2018.

[87] T. Back, *Evolutionary Algorithms in Theory and Practice: Evolution Strategies, Evolutionary Programming, Genetic Algorithms*. Oxford university press, 1996.

[88] W. Shi, J. Li, W. Xu, H. Zhou, N. Zhang, S. Zhang, and X. Shen, "Multiple Drone-Cell Deployment Analyses and Optimization in Drone Assisted Radio Access Networks," *IEEE Access*, vol. 6, 2018.

[89] E. Kalantari, H. Yanikomeroglu, and A. Yongacoglu, "On the Number and 3D Placement of Drone Base Stations in Wireless Cellular Networks," in *IEEE Vehicular Technology Conference (VTC-Fall)*, 2016.

[90] V. Roberge, M. Tarbouchi, and G. Labonté, "Fast Genetic Algorithm Path Planner for Fixed-Wing Military UAV Using GPU," *IEEE Transactions on Aerospace and Electronic Systems*, 2018.

[91] R. Poli, J. Kennedy, and T. Blackwell, "Particle Swarm Optimization," *Swarm intelligence*, vol. 1, no. 1, pp. 33–57, 2007.

[92] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Wireless Communication Using Unmanned Aerial Vehicles (UAVs): Optimal Transport Theory for Hover Time Optimization," *IEEE Transactions on Wireless Communications*, vol. 16, no. 12, pp. 8052–8066, 2017.

[93] B. Galkin, J. Kibilda, and L. A. DaSilva, "Deployment of UAV-mounted Access Points According to Spatial User Locations in Two-tier Cellular Networks," in *Wireless Days (WD)*, 2016, pp. 1–6.

[94] Z. Li, M. Kihl, Q. Lu, and J. A. Andersson, "Performance Overhead Comparison between Hypervisor and Container Based Virtualization," in *IEEE International Conference on Advanced Information Networking and Applications (AINA)*. IEEE, 2017, pp. 955–962.

[95] K.-t. Seo, H.-s. Hwang, I.-y. Moon, O.-y. Kwon, and B.-j. Kim, "Performance Comparison Analysis of Linux Container and Virtual Machine for Building Cloud," in *Advanced Science and Technology Letters*, vol. 66, 2014, pp. 105–111.

[96] D. Bernstein, "Containers and Cloud: From LXC to Docker to Kubernetes," *IEEE Cloud Computing*, vol. 1, no. 3, pp. 81–84, 2014.

[97] J. Dolezal, Z. Becvar, and T. Zeman, "Performance Evaluation of Computation Offloading from Mobile Device to the Edge of Mobile Network," in *IEEE Conference on Standards for Communications and Networking (CSCN)*. IEEE, 2016, pp. 1–7.

[98] M. V. Barbera, S. Kosta, A. Mei, V. C. Perta, and J. Stefa, "Mobile Offloading in the Wild: Findings and Lessons Learned Through a Real-life Experiment with a New Cloud-aware System," in *IEEE International Conference on Computer Communications (INFOCOM)*. IEEE, 2014, pp. 2355–2363.

[99] J. Oueis, E. C. Strinati, and S. Barbarossa, "Multi-parameter Decision Algorithm for Mobile Computation Offloading," in *IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2014, pp. 3005–3010.

[100] Y.-H. Kao, B. Krishnamachari, M.-R. Ra, and F. Bai, "Hermes: Latency Optimal Task Assignment for Resource-constrained Mobile Computing," *IEEE Transactions on Mobile Computing*, vol. 16, no. 11, pp. 3056–3069, 2017.

[101] M. A. Puente, Z. Becvar, M. Rohlik, F. Lobillo, and E. C. Strinati, "A Seamless Integration of Computationally-Enhanced Base Stations Into Mobile Networks Towards 5G," in *IEEE Vehicular Technology Conference*. IEEE, 2015, pp. 1–5.

[102] H. Wu, Q. Wang, and K. Wolter, "Methods of Cloud-path Selection for Offloading in Mobile Cloud Computing Systems," in *IEEE International Conference on Cloud Computing Technology and Science (CloudCom)*. IEEE, 2012, pp. 443–448.

[103] J. N. Al-Karaki and A. E. Kamal, "Routing Techniques in Wireless Sensor Networks: a Survey," *IEEE wireless communications*, vol. 11, no. 6, pp. 6–28, 2004.

[104] M. K. Marina and S. R. Das, "Ad hoc on-demand Multipath Distance Vector Routing," *Wireless communications and mobile computing*, vol. 6, no. 7, pp. 969–988, 2006.

[105] S. Kumar, S. Khimsara, K. Kambhatla, K. Girivanesh, J. D. Matyjas, and M. Medley, "Robust On-Demand Multipath Routing with Dynamic Path Upgrade for Delay-Sensitive Data over Ad Hoc Networks," *Journal of Computer Networks and Communications*, vol. 2013, 2013.

[106] S. Othmen, A. Belghith, F. Zarai, M. S. Obaidat, and L. Kamoun, "Power and Delay-aware Multi-path Routing Protocol for Ad hoc Networks," in *IEEE International Conference on Computer, Information and Telecommunication Systems (CITS)*. IEEE, 2014, pp. 1–6.

[107] S. Mueller, R. P. Tsang, and D. Ghosal, "Multipath Routing in Mobile Ad hoc Networks: Issues and Challenges," in *Performance tools and applications to networked systems*. Springer, 2004, pp. 209–234.

[108] S. Barbarossa, S. Sardellitti, and P. Di Lorenzo, "Computation Offloading for Mobile Cloud Computing Based on Wide Cross-layer Optimization," in *Future Network and Mobile Summit (FutureNetworkSummit)*, July 2013, pp. 1–10.

[109] 3GPP, "Mobility enhancements in heterogeneous networks," 3rd Generation Partnership Project (3GPP), TR, 2013.

[110] D. Xenakis, N. Passas, L. Merakos, and C. Verikoukis, "Mobility Management for Femtocells in LTE-Advanced: Key Aspects and Survey of Handover Decision Algorithms," *IEEE Communications Surveys Tutorials*, vol. 16, no. 1, pp. 64–91, First 2014.

[111] Z. Becvar, P. Roux, and P. Mach, "Fast Cell Selection with Efficient Active Set Management in OFDMA Networks with Femtocells," *EURASIP Journal on Wireless Communications and Networking*, vol. 2012, no. 1, 2012. [Online]. Available: http://dx.doi.org/10.1186/1687-1499-2012-292

[112] D. Xenakis, N. Passas, and C. Verikoukis, "An Energy-centric Handover Decision Algorithm for the Integrated LTE Macrocell-femtocell Network," *Computer Communications*, vol. 35, no. 14, pp. 1684–1694, 2012.

[113] J. Oueis, E. C. Strinati, and S. Barbarossa, "Small Cell Clustering for Efficient Distributed Cloud Computing," in *IEEE International Symposium on Personal, Indoor, and Mobile Radio Communication (PIMRC)*. IEEE, 2014, pp. 1474–1479.

[114] A. Machen, S. Wang, K. K. Leung, B. J. Ko, and T. Salonidis, "Live Service Migration in Mobile Edge Clouds," *IEEE Wireless Communications*, vol. 25, no. 1, pp. 140–147, 2018.

[115] S. Wang, R. Urgaonkar, T. He, M. Zafer, K. Chan, and K. Leung, "Mobility-Induced Service Migration in Mobile Micro-clouds," in *IEEE Military Communications Conference (MILCOM)*, Oct 2014, pp. 835–840.

[116] S. Sardellitti, G. Scutari, and S. Barbarossa, "Joint Optimization of Radio and Computational Resources for Multicell Mobile Cloud Computing," in *IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2014, pp. 354–358.

[117] ——, "Joint Optimization of Radio and Computational Resources for Multicell Mobile-edge Computing," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 1, no. 2, pp. 89–103, 2015.

[118] ——, "Distributed Joint Optimization of Radio and Computational Resources for Mobile Cloud Computing," in *IEEE International Conference on Cloud Networking (CloudNet)*. IEEE, 2014, pp. 211–216.

[119] C. Wang, F. R. Yu, C. Liang, Q. Chen, and L. Tang, "Joint Computation Offloading and Interference Management in Wireless Cellular Networks with Mobile Edge Computing," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 8, pp. 7432–7445, 2017.

[120] I. Farris, T. Taleb, M. Bagaa, and H. Flick, "Optimizing Service Replication for Mobile Delay-sensitive Applications in 5G Edge Network," in *IEEE International Conference on Communications (ICC)*. IEEE, 2017, pp. 1–6.

[121] O. Munoz, A. Pascual-Iserte, and J. Vidal, "Optimization of Radio and Computational Resources for Energy Efficiency in Latency-Constrained Application Offloading," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 10, pp. 4738–4755, 2015.

[122] C. You, K. Huang, H. Chae, and B.-H. Kim, "Energy-Efficient Resource Allocation for Mobile-Edge Computation Offloading," *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1397–1411, 2017.

[123] J. Wang, L. Zhao, J. Liu, and N. Kato, "Smart resource allocation for mobile edge computing: A deep reinforcement learning approach," *IEEE Transactions on Emerging Topics in Computing*, 2019.

[124] C. Wang, C. Liang, F. R. Yu, Q. Chen, and L. Tang, "Computation Offloading and Resource Allocation in Wireless Cellular Networks With Mobile Edge Computing," *IEEE Transactions on Wireless Communications*, vol. 16, no. 8, pp. 4924–4938, 2017.

[125] P. Mach and Z. Becvar, "Mobile Edge Computing: A Survey on Architecture and Computation Offloading," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 3, pp. 1628–1656, 2017.

[126] L. Pu, X. Chen, J. Xu, and X. Fu, "D2D Fogging: An Energy-efficient and Incentive-aware Task Offloading Framework via Network-assisted D2D Collaboration," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 12, pp. 3887–3901, 2016.

[127] A. Ksentini, T. Taleb, and M. Chen, "A Markov Decision Process-based Service Migration Procedure for Follow me Cloud," in *IEEE International Conference on Communications (ICC)*. IEEE, 2014, pp. 1350–1354.

[128] A. Beloglazov and R. Buyya, "Energy Efficient Resource Management in Virtualized Cloud Data Centers," in *IEEE/ACM international conference on cluster, cloud and grid computing*. IEEE Computer Society, 2010, pp. 826–831.

[129] S. Wang, R. Urgaonkar, M. Zafer, T. He, K. Chan, and K. K. Leung, "Dynamic Service Migration in Mobile Edge-clouds," in *IEEE IFIP Networking Conference (IFIP Networking)*. IEEE, 2015, pp. 1–9.

[130] S. Wang, R. Urgaonkar, T. He, K. Chan, M. Zafer, and K. K. Leung, "Dynamic Service Placement for Mobile Micro-clouds with Predicted Future Costs," *IEEE Transactions on Parallel and Distributed Systems*, vol. 28, no. 4, pp. 1002–1016, 2016.

[131] Z. Gao, Q. Jiao, K. Xiao, Q. Wang, Z. Mo, and Y. Yang, "Deep Reinforcement Learning Based Service Migration Strategy for Edge Computing," in *IEEE International Conference on Service-Oriented System Engineering (SOSE)*. IEEE, 2019, pp. 116–1165.

[132] J. Plachy, Z. Becvar, and E. C. Strinati, "Cross-layer Approach Enabling Communication of High Number of Devices in 5G Mobile Networks," in *IEEE International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*. IEEE, 2015, pp. 809–816.

[133] J. Plachy, Z. Becvar, S. M. Zafaruddin, and A. Leshem, "Nash Bargaining Solution for Cooperative Relaying Exploiting Energy Consumption," in *IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. IEEE, 2019, pp. 1–5.

[134] Z. Syed, Mohammad, J. Plachy, Z. Becvar, and A. Leshem, "Energy Consumption Performance of Opportunistic Device-to-Device Relaying Under Lognormal Shadowing," *accepted in IEEE Systems journal*, September.

[135] J. Plachy, Z. Becvar, P. Mach, R. Marik, and M. Vondra, "Joint Positioning of Flying Base Stations and Association of Users: Evolutionary-based Approach," *IEEE Access*, vol. 7, pp. 11 454–11 463, 2019.

[136] J. Nagle, "RFC 970: On Packet Switches with Infinite Storage," *Request For Comments*, 1985.

[137] A. Asadi and V. Mancuso, "Energy Efficient Opportunistic Uplink Packet Forwarding in Hybrid Wireless Networks," in *ACM Proceedings of the international conference on Future energy systems*. ACM, 2013, pp. 261–262.

[138] P. C. Sofotasios, M. K. Fikadu, S. Muhaidat, S. Freear, G. K. Karagiannidis, and M. Valkama, "Relay Selection Based Full-duplex Cooperative Systems Under Adaptive Transmission," *IEEE Wireless Communications Letters*, vol. 6, no. 5, pp. 602–605, 2017.

[139] R. Ma, N. Xia, H.-H. Chen, C.-Y. Chiu, and C.-S. Yang, "Mode Selection, Radio Resource Allocation, and Power Coordination in D2D Communications," *IEEE Wireless Communications*, vol. 24, no. 3, pp. 112–121, 2017.

[140] X. Lin, J. Andrews, A. Ghosh, and R. Ratasuk, "An Overview of 3GPP Device-to-device Proximity Services," *IEEE Communications Magazine*, vol. 52, no. 4, pp. 40–48, April 2014.

[141] Z. Zhou, S. Zhou, J. Cui, and S. Cui, "Energy-efficient Coperative Communication Based on Power Control and Selective Single-relay in Wireless Sensor Networks," *IEEE transactions on Wireless Communications*, vol. 7, no. 8, 2008.

[142] K. Cohen and A. Leshem, "A Time-Varying Opportunistic Approach to Lifetime Maximization of Wireless Sensor Networks," *IEEE Transactions on Signal Processing*, vol. 58, no. 10, pp. 5307–5319, Oct 2010.

[143] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, 10th ed. Academic, 1972.

[144] N. B. Mehta, J. Wu, A. F. Molisch, and J. Zhang, "Approximating a Sum of Random Variables with a Lognormal," *IEEE Transactions on Wireless Communications*, vol. 6, no. 7, pp. 2690–2699, July 2007.

[145] J. M. Meredith, "Spatial channel model for Multiple Input Multiple Output (MIMO) simulations," *3GPP TR 25.996, Release 6*, July 2017.

[146] H.-C. Yang and M.-S. Alouini, *Order Statistics in Wireless Communications: Diversity, Adaptation, and Scheduling in MIMO and OFDM Systems*. Cambridge University Press, 2011.

[147] F. D. Côté, I. N. Psaromiligkos, and W. J. Gross, "A Chernoff-type Lower Bound for the Gaussian Q-function," 2012.

[148] M. López-Benítez and F. Casadevall, "Versatile, Accurate, and Analytically Tractable Approximation for the Gaussian Q-Function," *IEEE Transactions on Communications*, vol. 59, no. 4, pp. 917–922, 2011.

[149] M. Lauridsen, L. Noël, T. B. Sørensen, and P. Mogensen, "An Empirical LTE Smartphone Power Model with a View to Energy Efficiency Evolution," *Intel® Technology Journal*, vol. 18, no. 1, pp. 172–193, 2014.

[150] 3GPP, "Study on LTE Device to Device Proximity Services; Radio Aspects," 3rd Generation Partnership Project (3GPP), TR, 2014.

[151] J. Meredith, "Study on Channel Model for Frequency Spectrum Above 6 GHz," *3GPP TR 38.900, Jun, Tech. Rep.*, 2016.

[152] Y. d. J. Bultitude and T. Rautiainen, "IST-4-027756 WINNER II D1. 1.2 V1. 2 WINNER II Channel Models," *EBITG, TUI, UOULU, CU/CRC, NOKIA, Tech. Rep., Tech. Rep*, 2007.

[153] M. Gen and R. Cheng, *Genetic Algorithms and Engineering Optimization*. John Wiley & Sons, 2000, vol. 7.

[154] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, and L. Dittmann, "Cloud RAN for Mobile Networks - A Technology Overview," *IEEE Communications surveys & tutorials*, vol. 17, no. 1, pp. 405–426, 2015.

[155] A. Al-Hourani, S. Kandeepan, and S. Lardner, "Optimal LAP Altitude for Maximum Coverage," *IEEE Wireless Communications Letters*, vol. 3, no. 6, 2014.

[156] I. Bor-Yaliniz, S. S. Szyszkowicz, and H. Yanikomeroglu, "Environment-aware Drone-base-station Placements in Modern Metropolitans," *IEEE Wireless Communications Letters*, vol. 7, no. 3, pp. 372–375, 2018.

[157] 3GPP, " Further advancements for E-UTRA physical layer aspects," 3rd Generation Partnership Project (3GPP), TS, 2010. [Online]. Available: http://www.3gpp.org/dynareport/36814.htm

[158] R. I. Bor-Yaliniz, A. El-Keyi, and H. Yanikomeroglu, "Efficient 3-D Placement of an Aerial Base Station in Next Generation Cellular Networks," in *IEEE International Conference on Communications (ICC)*, 2016.

[159] A. Fouda, A. S. Ibrahim, I. Guvenc, and M. Ghosh, pp. 1–5, 2018.

[160] R. C. Eberhart and Y. Shi, "Comparison Between Genetic Algorithms and Particle Swarm Optimization," in *International conference on evolutionary programming*. Springer, 1998, pp. 611–616.

[161] G. Bolch, S. Greiner, H. de Meer, and K. S. Trivedi, *Queueing Networks and Markov Chains*. John Wiley & Sons, Inc., 2006.

[162] Z. Becvar, J. Plachy, and P. Mach, "Path Selection Using Handover in Mobile Networks with Cloud-enabled Small Cells," in *IEEE Personal, Indoor, and Mobile Radio Communication (PIMRC)*, 2014, pp. 1480–1485.

[163] 3GPP, "Handover Procedures," 3rd Generation Partnership Project (3GPP), TS, 2014.

[164] O. Muñoz, A. P. Iserte, J. Vidal, and M. Molina, "Energy-latency Trade-off for Multiuser Wireless Computation Offloading," in *IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*. IEEE, 2014, pp. 29–33.

[165] 3GPP, "Policy and Charging Control (PCC); Reference points," 3rd Generation Partnership Project (3GPP), TS, 2015. [Online]. Available: http://www.3gpp.org/dynareport/29212.htm

[166] ——, " Radio Resource Control (RRC), Protocol specification," 3rd Generation Partnership Project (3GPP), TS, 2015. [Online]. Available: http://www.3gpp.org/dynareport/36331.htm

[167] J. Andrews, H. Claussen, M. Dohler, S. Rangan, and M. Reed, "Femtocells: Past, Present, and Future," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 3, pp. 497–508, April 2012.

[168] 3GPP, "Telecommunication management; Self-Organizing Networks (SON); Self-healing concepts and requirements," 3rd Generation Partnership Project (3GPP), TS, 2014. [Online]. Available: http://www.3gpp.org/dynareport/32541.htm

[169] N. Meghanathan, "Mobility Models for Wireless Ad hoc Networks," in *Proceeding of Research Experience for Undergraduates, Jackson State University.* Presented at the REU 2010, Jackson State University, 2010.

[170] G. Vivier, A. Agustin, J. Vidal, O. Muñoz, S. Barbarossa, L. Pescosolido *et al.*, "Scenario, Requirements and First Business Model Analysis," *Deliverable D21 of ICT-248891 STP FREEDOM project*, 2010.

[171] K. Ha, P. Pillai, W. Richter, Y. Abe, and M. Satyanarayanan, "Just-in-time Provisioning for Cyber Foraging," in *Proceeding of the 11th annual international conference on Mobile systems, applications, and services*, ser. MobiSys '13. New York, NY, USA: ACM, 2013, pp. 153–166. [Online]. Available: http://dx.doi.org/10.1145/2462456.2464451

[172] M. P. Wylie-Green and T. Svensson, "Throughput, Capacity, Handover and Latency Performance in a 3GPP LTE FDD Field Trial," in *IEEE Global Telecommunications Conference (GLOBECOM).* IEEE, 2010, pp. 1–6.

[173] S. Barbarossa, S. Sardellitti, and P. Di Lorenzo, "Communicating While Computing: Distributed Mobile Cloud Computing Over 5G Heterogeneous Networks," *IEEE Signal Processing Magazine*, vol. 31, no. 6, pp. 45–55, 2014.

[174] 3GPP, " User Equipment (UE) radio transmission and reception," 3rd Generation Partnership Project (3GPP), TS, 2015. [Online]. Available: http://www.3gpp.org/dynareport/36101.htm

[175] K. Wang, M. Shen, J. Cho, A. Banerjee, J. Van der Merwe, and K. Webb, "MobiScud: A Fast Moving Personal Cloud in the Mobile Network," in *Workshop on All Things Cellular: Operations, Applications and Challenges - AllThingsCellular.* ACM Press, 2015, pp. 19–24.

[176] C. Mehlführer, J. C. Ikuno, M. Simko, S. Schwarz, M. Wrulich, and M. Rupp, "The Vienna LTE Simulators-Enabling Reproducibility in Wireless Communications Research," *EURASIP EURASIP Journal on Advances in Signal Processing*, vol. 2011, p. 29, 2011.

[177] Huijun Li, S. Habibi, and G. Ascheid, "Handover Prediction for Long-term Window Scheduling Based on SINR Maps," in *IEEE International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*. IEEE, 2013, pp. 917–921.

[178] I. Eyal, F. Junqueira, and I. Keidar, "Thinner Clouds with Preallocation," in *HotCloud*. Citeseer, 2013.

[179] B. Toepelt. (2008) Atom Benchmarked: 4W Of Performance. [Online]. Available: https://www.tomshardware.com/reviews/Intel-Atom-Efficient,1981-17.html

[180] K. Ha, Z. Chen, W. Hu, W. Richter, P. Pillai, and M. Satyanarayanan, "Towards Wearable Cognitive Assistance," in *ACM Mobile systems, applications, and services*, 2014, pp. 68–81.

[181] A. Hadachi, O. Batrashev, A. Lind, G. Singer, and E. Vainikko, "Cell Phone Subscribers Mobility Prediction Using Enhanced Markov Chain Algorithm," in *IEEE Intelligent Vehicles Symposium*. IEEE, 2014, pp. 1049–1054.

[182] 3GPP, "Technical Specification Group Technical Specification Group Core Network and Terminals; Handover procedures (Release 13)," Technical specification, Tech. Rep., 2015.

[183] J. Oueis, E. Calvanese-Strinati, A. De Domenico, and S. Barbarossa, "On the Impact of Backhaul Network on Distributed Cloud Computing," in *IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*. IEEE, 2014, pp. 12–17.

[184] M. Jaber, M. A. Imran, R. Tafazolli, and A. Tukmanov, "5G Backhaul Challenges and Emerging Research Directions: A Survey," *IEEE Access*, vol. 4, pp. 1743–1766, 2016.

[185] J. Plachy, Z. Becvar, and E. C. Strinati, "Dynamic Resource Allocation Exploiting Mobility Prediction in Mobile Edge Computing," in *IEEE International Symposium on Personal, Indoor, and Mobile Radio Communication (PIMRC)*. IEEE, 2016.

[186] N. Samaan and A. Karmouch, "A Mobility Prediction Architecture Based on Contextual Knowledge and Spatial Conceptual Maps," *IEEE Transactions on Mobile Computing*, vol. 4, no. 6, pp. 537–551, 2005.

[187] A. Nadembega, A. Hafid, and T. Taleb, "A Destination and Mobility Path Prediction Scheme for Mobile Networks," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 6, pp. 2577–2590, 2015.

[188] OpenStreetMap Foundation, "OpenStreetMap," 2013.

[189] D. Stynes, K. N. Brown, and C. J. Sreenan, "A Probabilistic Approach to User Mobility Prediction for Wireless Services," in *International Wireless Communications and Mobile Computing Conference (IWCMC)*. IEEE, 2016, pp. 120–125.

[190] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[191] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*. John Wiley & Sons, 2015.

[192] J. Plachy, Z. Becvar, and P. Mach, "Path Selection Enabling User Mobility and Efficient Distribution of Data for Computation at the edge of Mobile Network," *Computer Networks*, vol. 108, pp. 357–370, 2016.

[193] Z. Becvar, M. Vondra, and P. Mach, "Dynamic Optimization of Neighbor Cell List for Femtocells," in *IEEE Vehicular Technology Conference (VTC Spring)*. IEEE, 2013, pp. 1–6.

[194] V. M. Nguyen, C. S. Chen, and L. Thomas, "Handover Measurement in Mobile Cellular Networks: Analysis and Applications to LTE," in *IEEE International Conference on Communications (ICC)*. IEEE, 2011, pp. 1–6.

[195] J. Plachy, Z. Becvar, E. C. Strinati, and N. di Pietro, "Dynamic Allocation of Computing and Communication Resources in Multi-Access Edge Computing for Mobile Users," *submitted to IEEE Transactions on Network and Service Management*, 2020.

[196] N. Nikaein, M. K. Marina, S. Manickam, A. Dawson, R. Knopp, and C. Bonnet, "OpenAirInterface: A Flexible Platform for 5G Research," *ACM SIGCOMM Computer Communication Review*, vol. 44, no. 5, pp. 33–38, 2014.

[197] O.-R. Alliance, "O-RAN WG1 Operations and Maintenance Architecture v02.00," O-RAN Alliance, Tech. Rep., 2020.

[198] S. Ahmadi, *LTE-Advanced: A Practical Systems Approach to Understanding 3GPP LTE Releases 10 and 11 Radio Access Technologies*, 1st ed. Academic Press, 2013.

[199] E. Tejaswi and B. Suresh, "Survey of power control schemes for lte uplink," *Int. Journal Computer Sci. and Inform. Technol*, vol. 10, p. 2, 2013.

[200] M. Lauridsen, A. Jensen, and P. Mogensen, "Reducing LTE Uplink Transmission Energy by Allocating Resources," in *IEEE Vehicular Technology Conference (VTC Fall)*, Sept 2011, pp. 1–5.

Appendix A

Energy consumption of mobile networks

One of the important benefits of exploiting the MEC is that the computation is done on the MEC host instead of the UE. Thus, energy for computation is not consumed at the UE, and its battery life time is prolonged. To show energy consumption of the state of the art algorithms and the proposed solutions, energy consumption model is defined in this section.

A.1 Uplink transmission power control

The energy consumption of the UE depends on its transmission power P^{Tx}. The transmission power in mobile networks is controlled by the BS, and is based on the received power S_{Rx} at the BS. At the BS there is a target received power $S_{\mathrm{Rx,target}}$, that should be met by appropriate P^{Tx}. Based on the channel quality, e.g., SINR, MCS for the transmission is selected. The MCS specifies modulation and code rate, thus, defining number of bits transmitted per symbol. The MCS follows SINR, as with higher SINR, higher MCS can be selected for a transmission. Transmission power control, as defined by the 3GPP in [31] depends on MCS and available bandwidth represented by RBs in LTE-A system. The SINR at receiver is proportional to the transmission power P^{Tx} at transmitter, path loss PL and interference from other cells. In LTE-A, the P^{Tx} required for selected MCS and given number of allocated RBs is defined, according to 3GPP [31, 198, 199], as follows:

$$P^{\mathrm{Tx}} = min\left(P_{\mathrm{MAX}}, P^0 + \alpha \cdot \mathrm{PL} + 10log_{10}\left(M\right) + \Delta_{\mathrm{TF}} + f\right) \tag{A.1}$$

where P^{MAX} is the maximum available transmission power (23 dBm for the UE class 3 [199]); $\alpha \in \{0, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$ corresponds to the path loss compensation factor, PL is the downlink path loss estimate, M stands for the number of assigned RBs, Δ_{TF} represents a closed loop UE specific parameter based on the applied MCS, f is a correction value (also referred to as a Transmission Power Control (TPC) command

[198, 199]), and parameter P^0 represents the power offset computed as:

$$P^0 = \alpha \cdot (SINR_0 + P^N) + (1 - \alpha) \cdot (P^{MAX} - 10log_{10}(M_0)) \tag{A.2}$$

where P^N is the noise power per RB, and M_0 defines the number of allocated RBs for the case when the UE would be transmitting with its maximal transmission power).

Parameters Δ_{TF} and f are used for dynamic adjustment of the transmission power to keep required SINR at the receiver. As we assume open loop power control, we can omit these parameters as indicated in [199]. The parameter α is set to 1 so the UE fully compensates the path loss. Under these assumptions (commonly considered in real mobile networks), we can simplify power offset to $P^0 = \alpha \cdot (SINR_0 + P_N)$ and then, (A.1) can be rewritten as:

$$P^{Tx} = min\{P^{MAX}, \alpha \cdot (SINR + P^N + PL) + 10log_{10}(M)\} \tag{A.3}$$

From the known transmission power, energy consumed by the transmission of data, with duration t^{UL} is calculated as:

$$E_R^{UL} = P^{Tx} \cdot t^{\mathrm{UL}} \tag{A.4}$$

Form the (A.3) it is seen, that the energy consumption depends on the SINR (MCS). To show how consumed energy changes with various SINR (MCS) values, an example of tradeoff between energy and duration of transmission of 100 kB using 10 RBs with PL=80 dB is shown in Figure A.1 This figure shows, that high energy is consumed if the transmission lasts a short time. Contrary, less energy is required if the transmission time is prolonged.

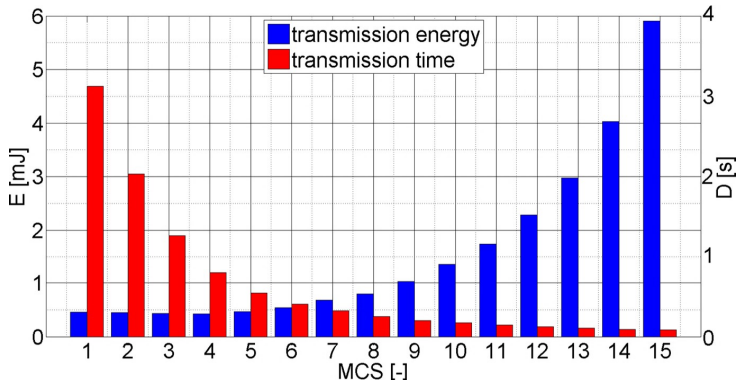


Figure A.1. Example of tradeoff between energy and time consumed by transmission of 100 kB using 10 RBs with path loss of 80 dB.

A.2 Empirical energy consumption model

A more advanced energy consumption model for the UE, based on a real measured UE data is described in [149, 200]. The energy consumption model consists of uplink and downlink energy consumption models, that are presented in this section. For the reception of computation results in downlink, the energy E_R^{DL} is derived based on knowledge of the power required for data reception (P_{Rx}). This power depends on the level of signal received by the UE (S_{Rx}) and data rate as specified in [174].

This energy consumption model considers the power consumption of the UE being turned on P_{ON}=853 mW, the uplink communication power P_{UL}, and the downlink communication power P_{DL}. In both uplink and downlink, the power consumption consists of the signal processing parts P_{TxBB} and P_{RxBB}, the radio parts P_{TxRF} and P_{RxRF}, and the consumption of the circuitry of communication parts P_{TxON} and P_{RxON}. Hence, the power consumed for the uplink communication (P_{UL}) is calculated as:

$$P_{UL} = P_{TxON} + P_{TxRF} + P_{TxBB} \text{ [mW]} \tag{A.5}$$

where $P_{TxON} = 29.9$ mW, $P_{TxBB} = 0.62$ mW, and P_{TxRF} is calculated as:

$$P_{TxRF} = \begin{cases} 0.78S_{Tx} + 23.6 & S_{Tx} \leq 0.2 \\ 17S_{Tx} + 45.4 & 0.2 < S_{Tx} < 11.4 \\ 5.9S_{Tx}^2 - 118S_{Tx} + 1195 & 11.4 < S_{Tx} \end{cases} \tag{A.6}$$

where S_{Tx} is the transmission power of the UE in dBm. Then, the energy consumed in the uplink for a transmission with a duration of t_{O} is calculated as:

$$E_{UL} = P_{UL} t_{\mathrm{O}} \tag{A.7}$$

The power consumed for downlink communication (P_{DL}) is calculated as:

$$P_{DL} = P_{RxON} + P_{RxRF} + P_{RxBB} \text{ [mW]} \tag{A.8}$$

where $P_{RxON} = 25.1$ mW, and P_{RxBB} is calculated as:

$$P_{RxBB} = 0.97R_{Rx} + 8.16 \text{ [mW]} \tag{A.9}$$

where R_{Rx} is the downlink throughput in Mbit/s, and P_{RxRF} is calculated as:

$$P_{RxRF} = \begin{cases} -0.04S_{Rx} + 24.8 & S_{Rx} \leq -52.5 \\ -0.11S_{Rx} + 7.86 & S_{Rx} > -52.5 \end{cases} \tag{A.10}$$

where S_{Rx} is the power received at the UE from the gNB in dBm. Similarly to the uplink,

energy consumed in the downlink with duration of t_{C}

$$E_{DL} = P_{DL} t_{\mathrm{C}} \tag{A.11}$$

The energy consumption of the UE is then calculated by multiplying the required power by the transmission time:

$$\begin{aligned} E = & E_{UL} + E_{DL} = P_{ON} \left(t_{\mathrm{O}} + t_{\mathrm{C}} + t_H \right) + \\ & + P_{DL} t_{\mathrm{C}} + P_{UL} t_{\mathrm{O}} \text{ [J]} \end{aligned} \tag{A.12}$$

Appendix B

Approximation of $I_2^{\mathrm{RELAY}}(N,\sigma)$

An approximation on $I_2^{\mathrm{RELAY}}(N,\sigma)$ is given as:

$$I_2^{\mathrm{RELAY}}(N,\sigma) \approx \sigma \sum_{k=0}^{N} \binom{N}{k} (-1)^k \Big(\frac{A(k)\gamma_{\max}}{\bar{\gamma}^2 + \bar{\gamma}\sigma\gamma_{\max}} + \frac{B(k)}{\bar{\gamma}} \log(1 + \frac{\sigma\gamma_{\max}}{\bar{\gamma}}) + C(k) \log|1 + \frac{\gamma_{\max}}{\alpha(k)}| + D(k) \log|1 + \frac{\gamma_{\max}}{\beta(k)}| \Big)$$

where

$$\begin{aligned} A(k) &= \tfrac{\sigma^2}{(\bar{\gamma}-\alpha(k)\sigma)(\bar{\gamma}-\beta(k)\sigma)}, B(k) = \tfrac{\sigma^2(\alpha(k)\sigma+\beta(k)\sigma-2\bar{\gamma})}{(\bar{\gamma}-\alpha(k)\sigma)^2(\bar{\gamma}-\beta(k)\sigma)^2} \\ C(k) &= \tfrac{1}{(\alpha(k)-\beta(k))(\alpha(k)\sigma-\bar{\gamma})^2}, D(k) = \tfrac{1}{(\alpha(k)-\beta(k))(\beta(k)\sigma-\bar{\gamma})^2} \\ \{\alpha(k),\beta(k)\} &= \big(-kq_2 \pm \sqrt{k^2q_2^2 - 4kq_1q_2 - 4kq_1} \big)/2kq_1 \\ q_1 &= -0.4920, q_2 = -0.2287, q_3 = -1.1893 \end{aligned} \tag{B.1}$$

Proof. To derive an approximate expression on $I_2^{\mathrm{RELAY}}(N,\sigma)$, we use an approximation on $Q(x) \approx \exp[-(q_1x^2+q_2x+q_3)]$ and $e^{-z} \leq \frac{1}{1+z}, \forall z \leq 0$ in (4.47) to represent the integral

$$I_2^{\mathrm{RELAY}}(N,\sigma) \approx \int_0^{\gamma_{\max}} \frac{\mathrm{d}x}{(x\sigma+\bar{\gamma})^2(1+k(q_1x^2+q_2x+q_3))} \tag{B.2}$$

where $\gamma_{\max} < \infty$ is chosen to avoid the divergence of the integral. The integration in (B.2) is derived in exact form as presented in (B.1). This completes the proof of Proposition 1. □

Appendix C

Scaling Law on Energy Consumption

We use $Q(0) = 1/2$ to get an upper bound on the integral $\mathcal{I}_1^{\text{RELAY}}(N, \sigma)$ in (4.42):

$$\mathcal{I}_1^{\text{RELAY}}(N, \sigma) \leq \frac{1}{2^N}(\frac{1}{\gamma_{\text{th}}} - \frac{1}{\bar{\gamma}}) \tag{C.1}$$

where the equality is achieved when $\gamma_{\text{th}} = \bar{\gamma}$. The integral $\mathcal{I}_2^{\text{RELAY}}(N, \sigma)$ in (4.42) can be decomposed:

$$\begin{aligned}\mathcal{I}_2^{\text{RELAY}}(N, \sigma) = &\int_0^{\delta_1} \frac{1}{(x\sigma + \bar{\gamma})^2}(1 - Q(x))^N \mathrm{d}x \\ &+ \int_{\delta_1}^{\delta_2} \frac{1}{(x\sigma + \bar{\gamma})^2}(1 - Q(x))^N \mathrm{d}x \\ &+ \cdots + \int_{\delta_M}^{\infty} \frac{1}{(x\sigma + \bar{\gamma})^2}(1 - Q(x))^N \mathrm{d}x\end{aligned} \tag{C.2}$$

where $\delta_i > \delta_{i-1} > 0$, $i = 1, 2, \cdots I$, where $I > 0$ is a positive integer. Since $Q(\delta_i) < Q(\delta_{i-1})$, we use the minimum of Q-function in each interval of integration to get an upper bound (C.2):

$$\begin{aligned}\mathcal{I}_2^{\text{RELAY}}(N, \sigma) \leq &(1 - Q(\delta_1))^N \frac{1}{\sigma}(\frac{1}{\bar{\gamma}} - \frac{1}{\sigma\delta_1 + \bar{\gamma}}) \\ &+ (1 - Q(\delta_2))^N \frac{1}{\sigma}(\frac{1}{\sigma\delta_1 + \bar{\gamma}} - \frac{1}{\sigma\delta_2 + \bar{\gamma}}) \\ &+ \cdots + \frac{1}{\sigma}(\frac{1}{\sigma\delta_I + \bar{\gamma}})\end{aligned} \tag{C.3}$$

We use $\delta_i = \sqrt{c_i \log(N)}$ where $0 \leq c_i \leq 1$, inequality $(1 - x)^N \leq \frac{1}{1+Nx}$, and a lower bound on Q-function $Q(x) \geq \kappa_2 e^{-x^2}$, where $\kappa_2 = 0.3885$ to bound $(1 - Q(\delta_i))^N$:

$$(1 - Q(\delta_i))^N \leq \frac{1}{1 + \kappa_2 N^{1-c_i}} \tag{C.4}$$

Using (C.4) in (C.3), we get

$$\mathcal{I}_2^{\text{RELAY}}(N,\sigma) \leq \frac{1}{\sigma}\Big[\frac{1}{\bar{\gamma}+\sigma\sqrt{c_I\log(N)}}+ \sum_{i=1}^{I-1}\Big(\frac{1}{1+\kappa N^{(1-c_i)}}\Big)\Big(\frac{1}{\bar{\gamma}+\sigma\sqrt{c_{i-1}\log(N)}}-\frac{1}{\bar{\gamma}+\sigma\sqrt{c_i\log(N)}}\Big)\Big] \tag{C.5}$$

where $c_0 = 0$. Using (C.1), (C.5) in (4.42), and neglecting negative terms, we get (4.47). When $N \to \infty$, the term involving $1/N$ becomes negligible, and we get the scaling law for energy consumption of Theorem 4.

Appendix D

List of Publications

Unless explicitly noted, the authorship is divided equally among the listed authors.

Publications related to the thesis

IF-journal papers

- J. Plachy, Z. Becvar, and P. Mach, "Path selection enabling user mobility and efficient distribution of data for computation at the edge of mobile network," *Computer Networks*, vol. 108, pp. 357-370, 2016.
  Citations except self–citations: 14 (WoS), 14 (Scopus), 21 (Google Scholar).

- J. Plachy [25 %], Z. Becvar [25 %], P. Mach [25 %], R. Marik [20 %], and M. Vondra [5 %], "Joint positioning of flying base stations and association of users: Evolutionary-based approach," *IEEE Access*, 7, pp.11454-11463, 2019.
  Citations except self–citations: 6 (WoS), 7 (Scopus), 16 (Google Scholar).

- J. Plachy, Z. Becvar, E. C. Strinati, N. di Pietro, "Dynamic Allocation of Computing and Communication Resources in Multi-Access Edge Computing for Mobile Users," *submitted to IEEE Transactions on Network and Service Management.*

- S. M. Zafaruddin, J. Plachy, Z. Becvar, A. Leshem, "Energy Consumption Performance of Opportunistic Device-to-Device Relaying Under Lognormal Shadowing". *accepted in IEEE Systems journal*, 2020.

Scopus and WoS-indexed conference papers

- J. Plachy [40 %], Z. Becvar [40 %], and E. C. Strinati [20 %], Cross-layer approach enabling communication of high number of devices in 5G mobile networks," *IEEE*

International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), IEEE, 2015.
Citations except self–citations: 4 (WoS), 5 (Scopus), 6 (Google Scholar).

- Z. Becvar, J. Plachy, and P. Mach, Path selection using handover in mobile networks with cloud-enabled small cells," *IEEE International Symposium on Personal, Indoor, and Mobile Radio Communication (PIMRC)*, 2014.
  Citations except self–citations: 12 (WoS), 12 (Scopus), 20 (Google Scholar).

- J. Plachy [40 %], Z. Becvar [40 %], and E. C. Strinati [20 %], Dynamic Resource Allocation Exploiting Mobility Prediction in Mobile Edge Computing," *IEEE International Symposium on Personal, Indoor, and Mobile Radio Communication (PIMRC)*, IEEE, 2016.
  Citations except self–citations: 46 (WoS), 56 (Scopus), 73 (Google Scholar).

- J. Plachy, Z. Becvar, S.M. Zafaruddin, and A. Leshem, "Nash Bargaining Solution for Cooperative Relaying Exploiting Energy Consumption". *IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, (pp. 1-5). IEEE, 2019.
  Citations except self–citations: 0 (WoS), 0 (Scopus),0 (Google Scholar).

Publications non-related to the thesis

Scopus and WoS-indexed conference papers

- Z. Becvar, M. Vondra, P. Mach, J. Plachy, D. Gesbert, "Performance of mobile networks with UAVs: Can flying base stations substitute ultra-dense small cells?." *European Wireless Conference (EW)* (pp. 1-7), 2017. **Best paper award**
  Citations except self–citations: 0 (WoS), 24 (Scopus), 39 (Google Scholar).

- Z. Becvar, P. Mach., J. Plachy and M.F.P. de Tudela, "Positioning of Flying Base Stations to Optimize Throughput and Energy Consumption of Mobile Devices." *IEEE Vehicular Technology Conference (VTC2019-Spring)*(pp. 1-7). IEEE, 2019.
  Citations except self–citations: 0 (WoS), 0 (Scopus), 0 (Google Scholar).

- J. Plachy, Z. Becvar, "Energy Efficient Positioning of Flying Base Stations via Coulomb's law." *accepted in IEEE Global Communications Conference (IEEE Globecom 2020) workshop on on Space-Ground Integrated Networks (SGINs)*, IEEE, 2020.
  Citations except self–citations: 0 (WoS), 0 (Scopus), 0 (Google Scholar).

Appendix E

List of Projects

Project no. SGS20/169/OHK3/3T/13, "Resource management based on machine learning for 6G mobile networks", funded by Czech Technical University in Prague, 01/2020-09/2020.

Project no. LTT20004 funded by Ministry of Education, Youth and Sports, "Cooperation with the International Research Centre in Area of Digital Communication Systems", 01/2020-09/2020.

Project "Mobile Edge Computing and Functional Splitting for Scheduling of Radio Resources", funded by Foxconn, 10/2018-09/2019.

Project no. LTT18007 funded by Ministry of Education, Youth and Sports, "Cooperation with the International Research Centre in Area of Communication Systems", 01/2018-12/2019.

Project no. P102/18/27023S funded by Czech Science Foundation, "Communication in Self-optimizing Mobile Networks with Drones", 01/2018-12/2020.

Project no. SGS17/184/OHK3/3T/13, "Flexible radio access for future mobile communications", funded by Czech Technical University in Prague, 01/2017-12/2019.

Project no. 8G15008, "Game theoretic aspects of wireless spectrum access", funded by Ministry of Education, Youth and Sports, 07/2016-06/2018.

Project Short research stay at EURECOM, France, funded by French government's scholarship programme, 06/2016-07/2016.

Project no. SGS13/199/OHK3/3T/13, "Mobile Self-Organizing Networks incorporating Small Cells", funded by Czech Technical University in Prague, 01/2015-12/2015.

Project no. ICT-318784, FP7 project TROPIC, founded by European Commission, 09/2012-04/2015.

Appendix F

Others

- J. Plachy, Z. Becvar, J. Dolezal, A. Ksentini, "Augmented Reality exploiting the Multi-Access Edge Computing in OpenAirInterface testbed", *Joint ETSI - OSA Workshop: Open Implementations and Standardization*, December 2018.