AMERICAN UNIVERSITY OF BEIRUT

User-Centric Strategies for Resource Management in Heterogeneous Wireless Networks with QoS Considerations

by

NADINE FAWAZ ABBAS

A dissertation
submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
to the Department of Electrical and Computer Engineering
of the Faculty of Engineering and Architecture
at the American University of Beirut

Beirut, Lebanon
April 2017

AMERICAN UNIVERSITY OF BEIRUT

User-Centric Strategies for Resource Management in Heterogeneous Wireless Networks with QoS Considerations

by

NADINE FAWAZ ABBAS

Approved by:

Dr. Hassan Artail, Professor — Chair of Committee
Electrical and Computer Engineering

Dr. Zaher Dawy, Professor — Co-Advisor
Electrical and Computer Engineering

Dr. Hazem Hajj, Associate Professor — Co-Advisor
Electrical and Computer Engineering

Dr. Youssef Nasser, Senior Lecturer — Member of Committee
Electrical and Computer Engineering

Dr. Mohsen Ghizani, Professor — Member of Committee

Electrical and Computer Engineering, University of Idaho

C/O

Dr. Sanaa Sharefeddine, Associate Professor — Member of Committee

Computer Science, Lebanese American University

Sanaa

Date of thesis defense: April 21, 2017

AMERICAN UNIVERSITY OF BEIRUT

THESIS, DISSERTATION, PROJECT RELEASE FORM

Student Name: Abbas Nadine Fawaz
Last First Middle

○ Master's Thesis ○ Master's Project ☑ Doctoral Dissertation

□ I authorize the American University of Beirut to: (a) reproduce hard or electronic copies of my thesis, dissertation, or project; (b) include such copies in the archives and digital repositories of the University; and (c) make freely available such copies to third parties for research or educational purposes.

☑ I authorize the American University of Beirut, to: (a) reproduce hard or electronic copies of it; (b) include such copies in the archives and digital repositories of the University; and (c) make freely available such copies to third parties for research or educational purposes after: **One ✓ year from the date of submission of my thesis, dissertation or project.**
Two ___ years from the date of submission of my thesis , dissertation or project.
Three ___ years from the date of submission of my thesis , dissertation or project.

Signature

May 5, 2017
Date

This form is signed when submitting the thesis, dissertation, or project to the University Libraries

Acknowledgements

Thank you God for all the benefits you have given me to accomplish my thesis, wishing that You bless my work and my future.

I am heartily thankful to my advisors Prof. Zaher Dawy and Prof. Hazem Hajj for all their assistance and support during this work and throughout my graduate studies. They encouraged and guided me from start to finish in developing a deep understanding of the subject. Their instructive comments and evaluation guided and challenged my thinking, substantially improving the thesis outcome.

I would like to acknowledge my thesis committee members, Prof. Hassan Artail, Prof. Mohsen Guizani, Prof. Sanaa Sharafeddine and Prof. Youssef Nasser for their dedicated time and constructive comments. I am also greatly thankful to the Electrical and Computer Engineering Department, AUB University Research Board (URB), and Lebanese National Council for Scientific Research (CNRS-L) PhD Scholarship Program. This work was made possible by NPRP Grant 7-1529-2-555 from the Qatar National Research Fund (a member of The Qatar Foundation); the statements made in this dissertation are solely my responsibility.

I would like to also thank Prof. Sanaa Sharafeddine and Mr. Karim Jahed from the Lebanese American University for making their Android heterogeneous networks simulator available to implement and evaluate various proposed algorithms.

The deepest gratitude goes towards my parents who provided me with moral support and bared my stress during the whole period. To my brother Mohammad who supported and encouraged me to keep going on, thank you. I hope I was able to make them proud.

I gratefully acknowledge my AUB family who supported me through the years and were aspiring to this moment. My sincere appreciation goes to Mrs. Rabab Abi Shakra for all the support and guidance she provided at the personal and academic levels. I would like to extend my gratitude to Prof. Joseph Costantine, Mr. Khaled Joujou and Mr. Ali Kaafarani for their continuous help and support.

Last but not least, no words can express my gratitude to my colleagues and friends. I am heartily thankful to my friend and sister Miss Reem Mahmoud who always stood by my side and helped me finding strengths in tough times. My deepest gratitude extends to Amal, Rajaa, Lama, Mohannad, Wissam, Saniya, Lynn, Tina and Diana, and the soon-to-be doctors Sireen, Maya, Mohammad, Carine, Abbas and Nour. I really cannot thank them enough and I pray that God preserves them for me.

An Abstract of the Thesis of

Nadine Fawaz Abbas for Doctor of Philosophy
Major: Electrical and Computer Engineering

Title: User-Centric Strategies for Resource Management in Heterogeneous Wireless Networks with QoS Considerations

Demand for mobile applications is increasing at an exponential rate which is loading existing wireless networks. The research community is currently actively involved in the design of new technologies that can enable massive device connections with the needed speeds and reliability. To this end, a major opportunity is to design solutions that facilitate the dynamic utilization and seamless operation of heterogeneous networks where devices can utilize multiple wireless interfaces simultaneously and cooperate with other devices in their vicinity. In this thesis, we propose and evaluate novel solutions to address emerging challenges related to the design of next generation heterogeneous wireless networks. Our research work is divided into two key objectives: the first objective aims at designing effective user-centric resource management techniques in cellular/WiFi heterogeneous networks with quality of experience considerations, and the second objective aims at optimizing traffic offloading in highly dense wireless networks using device-to-device cooperation, local caching, and planned channel allocation.

To achieve the first objective, we propose cellular/WiFi resource management strategies for a single-user scenario where a user can take advantage of the coexistence of multiple wireless interfaces to achieve performance gains. We first design a learning-based approach for network selection where a user utilizes one wireless interface at a time to achieve either minimum energy consumption, maximum throughput or energy efficiency based on user preferences. We then formulate the static traffic splitting problem, where a user utilizes both interfaces simultaneously, as a multi-objective optimization approach that captures the tradeoffs between throughput maximization on one hand and device battery energy minimization on the other hand. We then extend our work to address real-time traffic splitting decisions capturing the tradeoff between queue stability, energy consumption, and quality of experience for video streaming applications. The proposed strategies are evaluated using simulations and experimental test bed measurements under realistic operational conditions. In the second thesis objective, we propose multi-user traffic offloading strategies for dense wireless networks where a very high number of users in a given geographical area request simultaneously access to given data services, e.g., in a sports stadium or exhibition center. We formulate the problem as an optimization that aims at maximizing the number of served users while maintaining target quality of service using device-to-device cooperation, in-device caching, and intelligent channel allocation. Due to the complexity of the problem, we design sub-optimal hierarchical tree-based algorithms for real-time operation taking into account realistic constraints. We demonstrate their effectiveness by presenting performance results and analysis for a wide range of network scenarios.

Contents

Chapter 1

Introduction

According to Cisco Visual Networking Index, the mobile traffic will reach 49 Exabytes per month by 2021, where video streaming and downloads are expected to consist of more than 78% of all consumer Internet traffic [1]. To meet these tremendous traffic demands, the research community is currently actively involved in the design of the key components that will lead to the development of the 5G cellular technology, with the ability to accommodate massive connections and high load with ultra-fast speeds [2]. In addition, operators are looking for enhancing the quality of service (QoS) and satisfying users expectations and quality of experience (QoE). Along the same direction, the LTE-Advanced standard is evolving towards supporting advanced features such as small cell deployments and device-to-device (D2D) communications. The vision towards 5G is to allow the dynamic utilization of spectrum and multiple access technologies for the best delivery of services, including D2D communication, spectrum refarming and radio access infrastructure sharing. These lead to heterogeneous network deployments which require mobile devices to function under seamless operation over multiple wireless interfaces simultaneously, in order to boost throughput, capacity, coverage and quality of experience [3][4].

Heterogeneous networks (HetNets) are currently under intensive academic and industrial research due to their notable coverage and capacity gains as compared to conventional single-tier cellular networks. The idea behind HetNets is to overlay existing networks with additional infrastructure in the form of smaller low-power low-complexity access nodes. As shown in Figure 1.1, HetNets include macro cells served by base stations covering large coverage areas, and small cells served by low-power access nodes or mobile terminals for device-to-device cooperation. Small cells are primarily added to increase capacity in hotspots with high user demand and to fill in areas with weak coverage. Small cells include: (1) pico cells operated and managed by the network operator to provide coverage in hot spot sites such as malls, airports or stadiums, (2) femto cells such as WiFi access points, powered and connected by the end user, (3) relay nodes, which are low-power devices connected to the macro cell base station, (4) distributed antenna

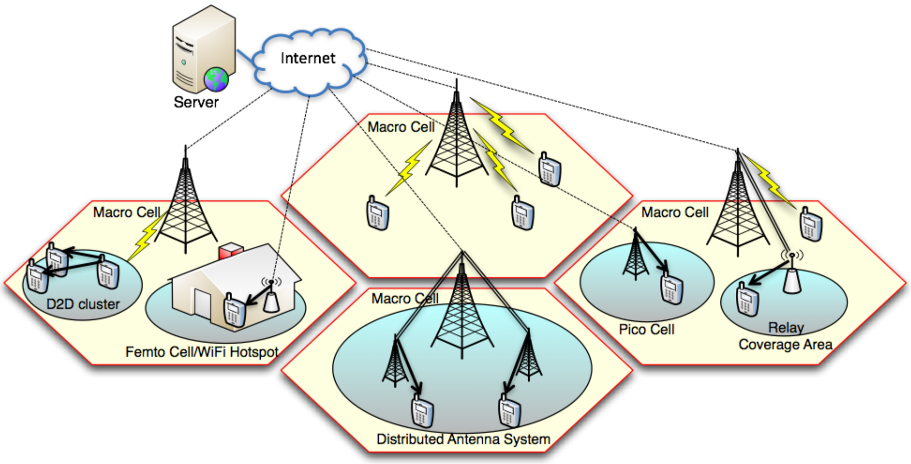


Figure 1.1: Heterogeneous networks formed by macro cells, small cells, pico cells and femto cells/WiFi hotspots.

system composed of spatially separated antenna nodes connected to a common source processing unit via fiber, and (5) D2D clusters, served by mobile terminals acting as cluster heads and transmitting data to other mobile terminals via short range connectivity. HetNets bring several technical challenges due to the presence of multi-radio access technologies (RATs), interference between non-orthogonal channels, as well as to the dense, decentralized, and dynamic deployment in licensed and unlicensed spectrum. The resources such as power and bandwidth should be allocated to maximize system performance such as throughput, capacity, coverage, fairness and energy efficiency. In addition, decisions need to adapt to the fast variation of the environment in terms of users' interface availability and channel conditions [5][6].

In this PhD thesis, we propose novel strategies for resource management with quality of service considerations to address emerging challenges related to the design of next generation heterogeneous wireless networks. Our research work is divided into two key research problems: the first aims at designing effective user-centric resource management techniques in cellular/WiFi heterogeneous networks with quality of experience considerations, as illustrated in Figure 1.2. We address static network selection and traffic splitting in HetNets to provide the user a balance between energy consumption and throughput based on user preferences. We then extend our work to address real-time traffic splitting decisions capturing the tradeoff between queue stability, energy consumption, and quality of experience for video streaming applications. We evaluate our proposed approaches using simulations and experimental test bed measurements under re-

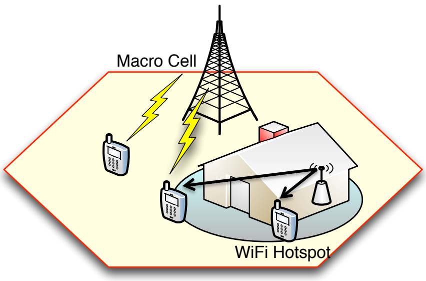


Figure 1.2: User-centric cellular/WiFi resource management download strategies in HetNets including network selection and traffic splitting.

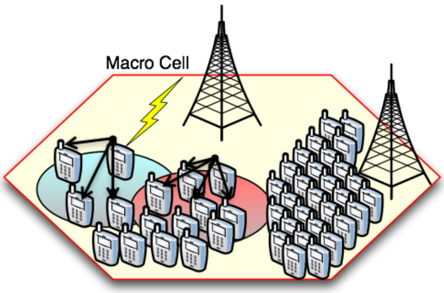


Figure 1.3: Resource management in dense D2D cooperative heterogeneous networks including traffic offloading and channel allocation.

alistic operational conditions. The second key research problem aims at optimizing resource management and traffic offloading in highly dense wireless networks using device-to-device cooperation, local caching, and planned channel allocation, as illustrated in Figure 1.3. We address multi-user resource management in dense wireless networks where a very high number of users request simultaneously access to given data services, such as in a sports stadium or exhibition center. We formulate the traffic offloading and channel allocation problems as optimization problems aiming at maximizing the system capacity, minimizing the use of cellular resources and offloading the traffic to D2D cooperation. We also present sub-optimal tree-based algorithms for traffic offloading with/without non-orthogonal channel allocation to provide near-optimal solutions in real network scenarios. We demonstrate the effectiveness of the proposed algorithms by presenting performance results and analysis for a wide range of network scenarios.

1.1 Thesis Objectives and Contributions

The PhD thesis work is divided into two main problems as follows: (1) user-centric cellular/WiFi resource management strategies considering network selection and traffic splitting, for data download and video streaming with QoE considerations, and (2) multi-user resource management including traffic offloading and non-orthogonal channel allocation in dense D2D cooperative heterogeneous networks.

1.1.1 User-Centric Cellular/WiFi Resource Management in HetNets

In this research direction, we address user-centric cellular/WiFi resource management strategies where users can utilize one wireless interface at a time, which is denoted as network selection, or utilize both interfaces simultaneously, which is denoted as traffic splitting, for data download and video streaming. A single user can take advantage of the coexistence of WiFi and cellular networks for enhancing its perceived quality of service and experience. We consider a heterogeneous network that is typically composed of areas covered via 3G/4G cellular macro cells and WiFi hotspots. The proposed methods decide on behalf of the user to use one link or multiple links simultaneously based on system parameters such as throughput, energy consumption, and quality of experience.

We solve two objectives in this direction: (1) static cellular/WiFi network selection and traffic splitting for data download, and (2) dynamic cellular/WiFi traffic splitting for video streaming with quality of experience considerations, and experimental evaluation using real test bed for cellular/WiFi traffic splitting under realistic operational conditions.

Objective 1: Static Cellular/WiFi Network Selection and Traffic Splitting for Data Download

In the first objective, we address static network selection and traffic splitting for data download where a user can make a decision only once at the start of the download request. In the first phase, we consider static network selection where a user can select between WiFi or cellular links to download the data. A learning-based approach for cellular/WiFi network selection is proposed. The best network is selected based on the channel conditions, throughput, download energy consumption of the device, type of application, and mobile device battery life as well as personal preferences. In the second part of the first objective, a static cellular/WiFi traffic splitting approach for data download is proposed. We formulate the traffic splitting problem as a multi-objective optimization approach that captures the tradeoffs between throughput maximization on one hand and device battery energy minimization on the other hand.

Objective 2: Dynamic Cellular/WiFi Traffic Splitting for Video Streaming with QoE Considerations

Traffic splitting in HetNets can provide higher throughput with a tradeoff cost in terms of energy consumption. However, there is a need to develop intelligent dynamic strategies for practical scenarios with focus on real-time streaming applications.As a second objective, we consider real-time traffic splitting decisions in cellular/WiFi HetNets to provide the user with high quality of experience while using video on demand streaming applications. In contrast to the literature, our proposed approach does not focus only on throughput and energy consumption, but also considers user quality of experience based on ITU-T P.1201 standard to better capture the video quality as perceived by the end user. We propose an optimized multi-objective traffic splitting approach based on Lyapunov drift-plus-penalty optimization utility functions aiming at achieving high quality end-user experience and minimizing energy consumption while stabilizing network queues. The proposed dynamic traffic splitting with delay-power-QoE balance approach determines in real-time how to intelligently split traffic between both wireless interfaces and decide dynamically on the amount of data to be sent over each wireless link. It is user-centric and provides autonomous real-time decision making at the user end without the need for any changes to the cellular/WiFi standards.

We evaluate our proposed approaches under realistic conditions using an experimental test bed. In current mobile devices, the traffic is offloaded directly to WiFi when WLAN is available. Smart mode option allowing auto-switching between WiFi and cellular data networks are now implemented in some smartphones such as iOS9 Iphones and Samsung Galaxy S5. This allows the device to switch from WiFi to cellular network only when WiFi network is not stable with poor connectivity. The test bed allows the user to use multiple wireless interfaces simultaneously which is not yet supported by current mobile devices to achieve performance gains.

The test bed is implemented using a modular approach which facilitates enhancements and extensions to implement and test various protocols, design alternatives, or intelligence options. Video streaming applications are deployed in the server acting as application service provider. The user will be downloading a video with specific size, duration and frame rate using a client application implemented on the Android platform using Java programming language. The application is installed at the user end, Android smartphones with 3G/WiFi network interfaces in our case, to make autonomous download strategies decisions taking channel conditions, energy consumption, quality of service and experience into consideration. The proposed approaches are implemented and tested under realistic conditions. To compare the performance of various strategies, we consider evaluating the queue size, the average throughput, total energy consumption, delay and QoE for different scenarios considering different channel conditions.

1.1.2 Multi-User Resource Management in Dense D2D Cooperative HetNets

In this research direction, we consider multi-user resource management in ultra dense device-to-device cooperative heterogeneous networks where users download a common content data such as file distribution and multimedia video on demand streaming.

Typical ultra-dense networks (UDNs) scenarios, presented in IMT-2020 (5G) Promotion Group [7], include dense urban areas, subways, metro stations, shopping malls, train stations, airports, open-air assembly and stadiums. The network density is the highest when a very high number of users request simultaneously a huge amount of data [8]. Common content distribution is nowadays considered one of the most emerging applications for mobile terminals, serving a large fraction of the Internet content including media files, documents, and multimedia streaming such as live and on-demand streaming. However, due to the huge growth of traffic demands and subscribers, macro network resources are becoming more scarce. To address the lack of network bandwidth, coverage and capacity, 3GPP considered new solutions using D2D cooperation [9]. In the cooperative network, the mobile terminals can cooperate to receive data using different wireless interfaces from the base stations or from other mobile terminals.

In addition to traffic offloading, the main challenge in ultra dense networks is channel allocation. The number of non-overlapping orthogonal channels is limited and the reuse of these channels becomes a must to serve the users. The use of non-orthogonal channels causes interference which decreases the achievable user throughput. Accordingly, considering channel allocation along with traffic offloading is needed to ensure more optimized network performance.

Objective 3: Traffic Offloading for Maximum User Capacity in Dense D2D Cooperative Networks

As a third objective, we address traffic offloading in dense D2D cooperative network where a large number of users request simultaneously common content data, assuming all the channels allocated are orthogonal. In our work, we consider data cache-enable mobile devices which act as content owners and distribute the common content data to other mobile terminals. Accordingly, in a cooperative environment, a mobile terminal can receive data either from a BS/AP over a long range wireless technology (such as WLAN, UMTS/HSPA, or LTE) or from other MT or content owner using a short range wireless technology (such as LTE-Direct, WiFi-Direct, Bluetooth or WiFi ad hoc mode).

We first formulate the traffic offloading problem as an optimization problem to find the optimal long range and short channel allocation constrained by the number of APs, LR and SR channels, number of content owners, users per cooperation cluster, and transmission rate. We solve the optimization problem using Advanced Interactive Multidimensional Modeling System (AIMMS) software and CPLEX as a solver considering a stadium topology as a case study to demonstrate the significant gains of optimized traffic offloading in ultra dense wireless networks.

The traffic offloading optimization problem can be shown to be NP-complete; optimal solutions may not be achievable in real-time ultra dense D2D cooperative networks. In addition, the optimization problem is holistic and considers all the existing users when providing optimal traffic offloading solutions. Accordingly, with the arrival of new users, the channel assignments and connections change to find optimal LR and SR channel allocation. This leads to frequent changes in the allocation of channels and the role of users as cluster heads, LR and SR users, which is not feasible in real-time networks. For these reasons, we propose a dynamic tree-based traffic offloading approach which assigns the users' connections consecutively based on a tree having BSs/APs and mobile terminals as nodes. We show that the proposed approach is able to provide near-optimal solutions with notably lower time complexity.

Objective 4: Non-Orthogonal Channel Allocation for Minimum Interference in Dense D2D Cooperative Networks

Serving a large number of users in dense HetNets is faced by limitations in the number of available orthogonal channels, especially that channel allocation takes into account different BSs/APs and cluster heads co-located in the same geographical area. Accordingly, careful channel allocation planning is needed to manage interference and enhance system performance. In our work, we address channel allocation to cluster heads based on the solution obtained from the optimal traffic offloading problem. We formulate the channel allocation as an optimization problem aiming at minimizing the reuse of the channels and maximizing the distance between non-orthogonal co-channel transmitters (cluster heads) to reduce the interference. We solve the optimization problem using AIMMS software and CPLEX as a solver. The results show that channel reuse may reduce the transmission service rate below service target rate. We then present possible solutions for allocating connections for the users affected by interference either to cluster heads or BSs/APs.

Allocating channels based on the solution provided by the traffic offloading assuming the channels are orthogonal may lead to a high user outage, which can be reduced by assigning the affected users to LR channels. Accordingly, providing sub-optimal solutions for traffic offloading with channel allocation considerations are needed to provide a balance between time complexity, user outage, number of LR channels and system performance. We propose a dynamic tree-based resource management approach which performs dynamic tree-based traffic offloading simultaneously with non-orthogonal channel allocation to cluster heads. The users' connections are assigned consecutively based on a tree having BSs/APs and mobile terminals as nodes. The channels are then assigned to cluster heads aiming at maximizing the distance between cluster heads using same channel. We show that the proposed approach is able to provide near-optimal solutions with low time complexity.

1.2 Organization of the Dissertation

The dissertation is organized as follows:

Chapter 2 presents a comprehensive literature review on existing approaches considering network selection, traffic splitting, traffic offloading with and without cache-enabled devices in dense device-to-device cooperative heterogeneous networks. For each approach, we highlight the proposed approaches, objectives and contributions. We compare and contrast the different approaches, as applicable. The various adopted assessment methods used in the evaluation of the proposed approaches are also discussed. Finally, we highlight the main contributions of this dissertation compared to the literature.

Chapter 3 presents first a learning-based user-centric approach for performing static network selection based on real-network implementations. This includes feature selection, training model development, network selection rules, experimental setup and network selection classification results. It also presents a multi-objective optimization approach for static traffic splitting, capturing the tradeoff between throughput maximization and energy consumption minimization.

Chapter 4 highlights the benefits of traffic splitting for real-time video streaming applications. It presents the proposed dynamic Lyapunov-based traffic splitting approach providing the user with a balance between high quality of experience, low energy consumption and delay while using video on demand streaming applications. It describes the Lyapunov drift-plus-penalty optimization formulation and the cost function expressed in terms of power consumption and QoE. The test bed implementation and setup is presented as well as performance results under realistic conditions.

Chapter 5 addresses traffic offloading to maximize the user capacity in dense D2D cooperative heterogeneous networks. It first presents the optimization formulation for resource management where non-orthogonal channel allocation is simultaneously considered with traffic offloading. The complexity of the problem is studied and reduced to address first optimal traffic offloading problem assuming all the channels are orthogonal. We then present a sub-optimal traffic offloading approach to provide real-time solutions in ultra dense D2D cooperative networks. The performance results of the proposed dynamic tree-based traffic offloading algorithm and complexity analysis are presented. To evaluate our solution, we focus on a stadium topology to demonstrate the significant gains of optimized traffic offloading in conventional and D2D ultra dense wireless networks with/without cache-enabled devices.

Chapter 6 presents optimal non-orthogonal channel allocation for minimum interference to cluster heads based on solutions obtained from the optimal traffic offloading problem. We then present possible solutions for allocating connections to users affected by interference either to cluster heads or BSs/APs. We finally propose a dynamic tree-based resource management approach where traffic offloading is simultaneously considered with non-orthogonal channel allocation. Results are also presented to evaluate the system performance.

Finally, Chapter 7 summarizes the main contributions of this dissertation. Moreover, it identifies the shortcomings of state-of-the-art wireless technologies and highlights the needs for enhancing technical standards in order to facilitate the implementation of the proposed concepts and algorithms in existing wireless networks. Chapter 7 also presents open research problems that need further investigation.

Chapter 2

Literature Review

This chapter presents a general overview on resource management in heterogeneous networks including content distribution with device-to-device cooperation. Section 2.1 surveys existing network selection and traffic offloading schemes in heterogeneous networks. Section 2.2 discusses various approaches for resource management in ultra dense networks with/without cache-enabled device-to-device cooperation. At the end of every section, a summary is presented to highlight the limitations of the studies conducted in the literature and show our work and contributions.

2.1 Network Selection and Traffic Splitting in HetNets

The research on cellular/WiFi heterogeneous networks can be divided between approaches focused on network selection, where users can utilize one wireless interface at a time, and approached focused on traffic splitting, where users utilize both interfaces simultaneously.

2.1.1 Network Selection in Heterogeneous Networks

The authors in the following research works focused on proposing approaches for the user to select one network in HetNets where different networks co-exist such as WiFi and cellular networks. The best network can be defined based on the coverage, cost, bandwidth and application QoS requirements, capacity, as well as personal preferences [10]. The approaches can be divided between user-centric and network-centric network selection approaches, handover and load offloading schemes.

In [11–13], the authors proposed user-centric distributed network selection approaches where mobile users strive to improve their performances on their own. The authors in [11] proposed a user-centric network selection method that allows decision making based on the user context and preferences. The authors in [12] presented a network selection approach for energy-delay tradeoff using the Lyapunov optimization framework for video upload considering queue stability, power consumption and throughput. In [13], a fuzzy logic based network selection approach was proposed considering signal strength, network load and mobile movement speed.

In [14–18], the authors addressed network-centric centralized network selection approaches where the network performs resource allocation maximizing global network utility while satisfying mobile users requirements. The authors in [14] used semi-Markov decision process and Q-learning to perform network-assisted network selection satisfying operator objectives while maximizing user preferences and requirements. In [15], the authors proposed a joint network selection and resource allocation for multicast in HetNets aiming at minimizing the overall bandwidth cost. In [16], the authors proposed a Q-learning and reinforcement learning-based approach for radio access technologies selection non-cooperative game where the throughput is the main objective function. The authors in [17] used particle swarm optimization and a modified version of the genetic algorithm to solve network selection and channel allocation providing users with a target quality of service at a low price, subject to the interference constraints. In [18], the authors formulated network selection as a global centralized non-linear optimization problem aiming at minimizing total users' cost service time and proposed heuristic distributed approaches and learning-based non-cooperative game to reduce the complexity and approximate the optimal solutions.

In [19–21], handover schemes were proposed. The authors in [19] addressed handover and dynamic adaptation of mobile devices while investigating the performance of the network in handling decision-making requests and network response time as the user satisfaction metric. In [20], the authors proposed a cooperative vertical handoff decision algorithm based on game theory to achieve the load balancing and meet the quality of service requirements of various applications. In [21], the authors proposed a handover scheme based on a utility function providing balance between energy, throughput and cost based on the user preferences.

In [22–24], the authors considered traffic offloading in HetNets from the cellular network to WiFi to balance the load, accommodate new requests and reduce network congestion. In [22], the network selection is addressed at the network side by offloading users from the cellular network to WiFi for load balancing using joint resource partitioning. The authors in [23] proposed a heuristic approach to dynamically adjust the workload of heterogeneous base stations to accommodate new requests. In [24], the authors aimed to reduce network congestion by proposing an incentive mechanism to allow network operators to optimally reward users to participate in delayed WiFi offloading. The authors in [25] addressed power and admission control in small cells deployment aiming at maximizing user admission, network spectral and energy efficiency while satisfying users minimum rate requirements.

Despite the enhancements provided by the intelligent network selection approaches, the performance in terms of system capacity, bandwidth and user quality of experience is still limited. Thus, utilizing simultaneously multiple wireless interfaces in the devices via dynamic traffic splitting mechanisms are proposed to achieve more optimized performance.

2.1.2 Traffic Splitting in Heterogeneous Networks

The authors in the following research works considered traffic splitting in HetNets where users can simultaneously use multiple links to download data. In [26–29], the user receives data over both interfaces consecutively according to a given ratio. The authors in [26] aimed to split the traffic periodically according to the ratio of the time required to send the data using one link over the sum of time required to send it via both links. In [27], the authors worked on minimizing the time required to send the data over WiMAX and WiFi to determine the traffic splitting ratio. In [28], Yang et al. proposed a traffic splitting approach based on a split ratio dynamically adjusted based on the network channel quality and load to enhance throughput. The authors in [29] presented a multipath packet transmission scheme that reduces packet reordering, improves throughput and maximizes utilization of the links. The faster link is assigned more packets than the slower one.

In [30–33], the authors aimed to split the traffic based on the service characteristics. The control or important information such as base layer in video are sent over 3G while others are sent via WiFi. In [30], the control or important information such as base layer in video are sent over 3G while others are sent via WiFi. The authors in [31] proposed layered video streaming allocation based on traffic characteristics to increase system capacity. The authors in [32] proposed a delay tolerant approach in heterogeneous networks where the user sends a request via 3G to the base station which replies by forwarding the requested content via 3G or WiFi based on links' availability. In [33], the I-frames, containing the full information of the video, are sent over 3G with a guaranteed level of quality of service while other frames are sent via WiFi.

The authors in [34] proposed a centralized multi-RAT bandwidth aggregation where LTE and WiFi networks are used to transfer different services simultaneously taking into consideration networks congestion. The authors in [35] proposed a traffic splitting approach performed at the network level, through jointly optimizing traffic control and radio resource allocation of multiple radio access networks. The authors in [36] and [37] considered uplink traffic splitting and scheduling. The authors in [36] proposed a packet scheduling algorithm based on parallel aggregation of radio nodes transmission schemes to improve the delay performance. The authors in [37] proposed resource allocation for uplink traffic splitting while considering a particular aspect of user QoE, where the QoE metric focuses on: (1) link reward function reflecting the achieved throughput as quality of service, and (2) resource cost function representing the cost required to use the allocated resources per unit bandwidth. They aimed at illustrating the trade-off between the link throughput and associated cost without considering actual user experience as perceived by the user end.

2.1.3 Summary and Contributions

Previous network selection and traffic splitting approaches considered load balancing, system capacity enhancement, bandwidth allocation, throughput maximization and power consumption reduction. However, they did not consider simultaneously user quality of experience. Some work addressed system performance enhancement, and based their results on simulations or arbitrarily generated values for factors affecting network selection decisions such as signal quality, network load and achievable data rate. Some proposed approached were not dynamic, the decision is static and made only once. Traffic splitting decisions in some works were based on traffic characteristics, and may require network assistance.

In this thesis, we first present a learning-based approach for static network selection. In contrast to literature, the model considers the features that affect the selection decision known by the user: availability of the networks, signal strength reflecting the channel quality, data size, battery life, speed of the user, location, and type of application. We present an approach for building training data as a basis for machine learning of network selection and then develop decision-tree classification model for network selection that provides the user either the highest quality of service, lowest energy consumption or highest energy efficiency based on pre-defined rules.

We then present optimal traffic splitting in HetNets to guarantee a balance between energy consumption and throughput based on the user's needs in terms of application service requirements and mobile device battery life. Moreover, experimental measurements are used to determine values for key parameters in order to evaluate the proposed traffic splitting approach under realistic network conditions.

We also propose a Lyapunov-based multi-objective dynamic traffic splitting approach. In contrast to the literature, our approach does not focus only on throughput and energy consumption, but also considers user quality of experience based on ITU-T P.1201 standard to better capture the video quality as perceived by the end user. We evaluate the proposed approach under realistic operational conditions using our own test bed. The proposed approaches are implemented as an Android application that functions in the background at the user side without any intervention from the network or the server, and without performing any changes to the cellular/WiFi standards.

2.2 Multi-User Resource Management in Ultra Dense D2D Cooperative HetNets

Typical UDNs scenarios include dense urban areas, subways, metro stations, shopping malls, train stations, airports, open-air assembly and stadiums. The network density is the highest when a very high number of users request simultaneously a huge amount of data. Common content distribution is nowadays considered one of the most emerging applications for mobile terminals such as live streaming and video on-demand. To meet the tremendous traffic demands faced by the lack of macro resources, device-to-device cooperation is proposed to increase system coverage and capacity.

In this section, we first present existing literature addressing device-to-device cooperation in heterogeneous networks for common content distribution. Research on cache-enabled devices in D2D cooperation and resource management in ultra dense networks is presented.

2.2.1 D2D Cooperation in Heterogeneous Networks

Existing literature on cooperative common content distribution over wireless networks aims at one or more of several objectives such as: increasing the network throughput, decreasing monetary cost, decreasing the file download time, and decreasing the energy consumption.

The authors in [38–45] focused on increasing the network throughput. In [38–41], the authors addressed resource blocks and channel allocation aiming at maximizing throughput in D2D cooperative networks. In [38], the authors investigated the network throughput achieved by both spatial diversity and spatial frequency reuse in a wireless ad-hoc network. The authors in [39] addressed resource blocks allocation for D2D pairs using Markov approximation and matching-game approaches. In [40], the authors proposed resource allocation approach to coordinate the interference and maximize the system sum-rate when several D2D pairs communicate by reusing the resources of a cellular user. The authors in [41] presents a mesh adaptive search algorithm for solving the joint admission control, mode selection and power allocation problem in device to device communication, aiming at maximizing system throughput. The authors in [42–45] traffic offloading and resource allocation in D2D cooperative networks. The authors in [42] proposed grouping vehicles into collaborative clusters and selecting the best subcarriers for LTE transmission to enhance the received video quality, quality of service and throughput. Interference-aware multi-hop cooperative routing, dynamic available channel assignment, and relay selection are used to improve the throughput in [43]. The authors in [44] proposed a load balancing approach where D2D users can multiplex the spectrum allocated to a number of cellular users to maximize total throughput. In [45], the authors proposed region-based clustering in vehicular ad-hoc networks and non-overlapping radio channel allocation while limiting the number of vehicles in each service region unit in order to reduce the contentions for radio channels and increase throughput.

In [46] and [47], cost effective solutions are proposed in wireless networks. The authors in [46] proposed distortion controlled streaming video services providing near-optimal cost considering transmission cost per byte of networks. In [47], the authors presented a monetary cost-effective collaborative streaming among mobile devices providing high performance in terms of delay and cost fairness.

Further more, many existing works [48–60] concentrate on energy efficiency and power consumption reduction. The authors in [48] compared the average total transmit power of cooperative transmission with the average transmit power of conventional communication and showed savings in terms of power. In [49], the authors introduced an energy consumption factor based multi hop routing that made the network topology distributed more uniform, prolong the network lifetime, and enhance its performance. The authors in [50–56] aimed at proposing traffic offloading and connection assignment approaches in D2D cooperative networks to reduce energy consumption. The authors in [50] proposed an energy efficient nearest-neighbor cooperation communication scheme by exploiting the short-range communication between a MS and its nearest neighbor to collaborate on their uplink transmissions. The authors in [51] proposed a Nash bargaining solution for energy-efficient content distribution in mobile-to-mobile cooperation wireless networks. The authors proposed an optimal mobile terminal grouping in [52] for cooperative common content distribution for minimizing energy consumption. In [53], a distributed algorithm for coalition formation is proposed where terminals can cooperate for sharing content while minimizing the networks energy consumption. The authors addressed in [54] the problem of offloading the cellular network while distributing common content to a group of mobile devices that cooperate during the download process by forming device-to-device communication networks with fairness constraints. In [55], the required number of cellular channels is reduced subject to energy consumption constraints at the mobile terminals side. The authors in [56] proposed energy efficient application-aware multimedia delivery solutions including quality adaptation and missing content retrieval in cooperative HetNets where a device can download content from the neighboring device with the same interest on the content and providing lower energy consumption. In [57–60], the authors addressed power allocation and control maximizing energy efficiency. In [57], the author proposed a game-theoretic power cooperative control algorithm to minimize the total power consumption in a cooperative communication network that transmits information from multiple sources to a destination via multiple relays to save energy and improve communication performance. In [58], an optimum strategy of power and time allocations were proposed to minimize the outage probability of the ideal cooperative protocol. To maximize the energy efficiency of the cooperative multicast communication, the authors in[59] proposed a probability-based relay selection and power control methods. The authors in [60] proposed finding the optimal power control, relay assignment, and channel allocation such that all transmission rate requirements are satisfied and the total energy consumption is minimized.

2.2.2 Cache-Enabled D2D Cooperation in Heterogeneous Networks

Content sharing through device-to-device communications has been proven to be a promising method to offload the traffic of base stations. If some user devices have cached a few popular on-demand contents, other interested neighbor mobile terminals can reuse these contents through D2D communications, in which the contents are directly transmitted to the mobile terminals from the content owners. Hereby, the base station would only transmit contents which are not locally available instead of transmitting the same popular contents multiple times. Therefore, the traffic of the BSs is significantly offloaded using cache-enabled D2D cooperation.

Caching improves spectrum utilization, increases network throughput, and reduces average access delay for mobile terminals. However, in reality, users are selfish and only care about their own preferences. On the other hand, the base station aims at minimizing its traffic load and transmission cost by offloading to D2D communication. To motivate content owners, the authors in [61–63] proposed incentive mechanisms rewarding D2D users for participating in D2D content sharing. In [61], the authors proposed a non-monetary energy-aware incentive mechanism and a non-transferable utility coalition formation game for user grouping. The authors in [62] and [63] have modeled the interaction between the BS and end-users by Stackelberg game models. In [62], the authors proposed a game theoretic approach to content trading in proactive wireless networks to maximize the profit for wireless network carrier and minimize payment for end-users. The authors in [63] introduced an Stackelberg game based incentive mechanism to encourage content sharing among mobile terminals by determining reward policies minimizing the BS total cost and caching policies maximizing mobile terminals utility.

The work in [64–77] addressed cache-enabled D2D cooperation to improve system performance in heterogeneous networks. In [64], the authors presented a comprehensive analytical framework to show that cache-enabled D2D communication provide higher performance as the requesting users move away from the BS and most popular files are requested. In [65–69], the authors proposed cashing policies and scheduling minimizing delay [65], maximizing the content-related energy efficiency [66], maximizing successful offloading probability [67], and enhancing the success rate of content fetching [68]. The authors in [69] aimed at maximizing traffic offloading while reducing energy costs for the D2D network with optimized proactive caching policy and transmit power.

The authors [70–77] proposed cooperation strategies including traffic offloading in caching-enabled D2D cooperative networks. The authors in [70] modeled the content caching problem at network edge nodes as a Markov decision process and proposed a distributed cache replacement strategy based on Q-learning to minimize the transmission cost. The authors in [71–74] aimed at addressing traffic offloading and user assignments maximizing system throughput. In [71], Chen et al. proposed cooperation strategy for cache-enabled D2D communications to manage the interference among D2D links, maximize the average network throughput by jointly optimizing the cluster size and bandwidth allocation satisfying minimum user average data rate. The authors in [72] address traffic offloading to maximize the user throughput and the system capacity in heterogeneous networks where users retrieve, in priority, contents from neighbors caches using single-hop and two-hop D2D communication. In [73], the authors proposed a content caching and replacement scheme to minimize content retrieving delay considering the limited storage capacity of mobile terminals, content popularity, and content access process in the network. In [74], the authors aimed first at finding the optimal cluster size maximizing network throughput while ensuring user fairness, and then optimizing bandwidth partition to maximize the average network throughput under the constraint on average user rate. Park et al. in [75] aimed at alleviating network congestion by proposing content management protocols to determine the amount of traffic to be offloaded to the D2D network among the cellular operator, the D2D servers, and the D2D clients. The authors in [76] proposed a greedy intra-cluster cache scheme by combine user clustering and file clustering according to file preferences in designing D2D caching schemes. In [77], Jiang et al. proposed an interference-aware communication model including selective caching and sender-receiver matching to maximize traffic.

2.2.3 Ultra Dense Heterogeneous Networks

Due to the exploding traffic demands with the ubiquitous anticipated spread of 5G and Internet of Things, research has been active to devise mechanisms for meeting these demands while maintaining high quality of service. Ultra-dense networks has been widely considered as one of the key scenarios in 5G networks with the need to accommodate massive connections with ultra-fast speeds. Large number of small cells and access points are deployed in ultra dense network to improve the network capacity by offloading the tremendous macro cell traffic, balancing network loads, and reducing congestion. UDNs consist of macro cells, small cells, D2D links and relays, which collectively increase the complexity of the network environment and leads to high interference due to large frequency reuse factor. There have been extensive ongoing researches on ultra dense networks considering user association, interference management, energy efficiency, spectrum sharing, resource management, scheduling, backhauling, propagation modeling, and the economics of UDN deployment [78].

In this section, we present existing literature on dense heterogeneous networks where a large number of users request common content data simultaneously. In [79–88], dense cell deployment networks without D2D cooperation are considered to enhance throughput and energy efficiency. The authors in [79] evaluated the performance of UDN in terms of throughput, spectral and energy efficiency and determined their relationship with BS density. In [80], the authors used power control and user scheduling to optimize energy efficiency of UDNs. In [81], a subband allocation scheme based on the graph clustering theory is proposed to minimize the subband handoff rate in two-tier ultra-dense deployed heterogeneous networks. The authors in [82] proposed a combined access selection and load offloading from macro base stations to fixed virtual node base stations improving the network energy efficiency.

Stadiums are considered one of the most challenging ultra dense networks. Stadium is a place or venue for outdoor sports, concerts, or other events and consists of a field or stage either partly or completely surrounded by a tiered structure designed to allow spectators to view the event. In [83], Jevremovic presented the challenges designing an in-building wireless network for stadiums, such as: stadium sectorization planning, sector overlap management, macro interference management, capacity dimensioning, defining radio frequency coverage area, and macro handoff management. In [84–89], the authors addressed ultra dense stadium network model. The authors in [84] and[85] used machine learning including reinforcement learning, bayesian network model, Q-learning, for spectrum sensing and subchannels allocations reducing blocking, re-transmission and interruption probabilities in a stadium network model. The authors in [86] analyzed the capacity and coverage of indoor stadiums considering scattering and reflections of signals from human bodies. The authors in [87] investigated the spectral efficiency per stadium seating area for different deployment scenarios, technologies such as WiFi 802.11a, 802.11g and LTE and reuse factors. In [88], the authors proposed a dynamic spectrum resource utilization with multiple carrier deployment in dense stadium maximizing system throughput subject to mobile terminals rate demands.

In [90–93], dense cell deployment networks with D2D cooperation are considered. In [90], the authors analyzed the energy consumption of single hop and multi hop in cooperative dense ad-hoc networks. The authors in [91] proposed a hierarchical architecture for channel allocation aiming at minimizing the latency. In [92], an online learning algorithm for spectrum allocation is proposed to increase throughput, spectral efficiency, fairness, and reduce outage ratio. The authors in [93] proposed clustering, power control, frequency assignment and transmission scheduling techniques in dense wireless networks where WiFi-Direct is used for D2D communication.

The authors in [94,95] considered caching in ultra dense network. The authors in [94] considered cache-enabled base stations. They first explored a big data enabled platform which parallelizes the computation of content popularity via machine learning tools and cache contents at the base stations, and then studied various caching scenarios to assess performance gains for 5G wireless networks. In [95], Song et al. considered large-scale cache-enabled mobile helpers with D2D cooperation in dense heterogeneous networks. They proposed a contention based multimedia delivery protocol and a content caching strategy where the most popular file is cached in the library to maximize the successful content delivery probability.

2.2.4 Summary and Contributions

To this extent, the research on traffic offloading and channel allocation in ultra dense cache-enabled D2D cooperative networks is still limited. Existing work addressed traffic offloading to reduce the congestion on the macro cells, however, the main focus was on enhancing system throughput, increasing energy efficiency and minimizing the cost of transmission. In addition, the system model and results were based on limited number of users.

In contrast to the literature, our system model addresses simultaneously traffic offloading and channel allocation in ultra dense cache-enabled D2D cooperative heterogeneous networks. We present optimal solutions for resource management including traffic offloading and non-orthogonal channel allocation aiming at maximizing the system capacity while maintaining user target quality. We also present sub-optimal hierarchical tree-based algorithms to perform dynamic traffic offloading and channel allocation. In our work, we focus on a stadium topology considering thousands of users to demonstrate the significant gains of optimized traffic offloading in conventional and D2D ultra dense wireless networks with/without cache-enabled devices. Performance results and complexity analysis are presented to show that the proposed approaches are able to provide near-optimal solutions with notably lower time complexity.

Chapter 3

Static WiFi/Cellular Network Selection and Traffic Splitting for Data Download

In this chapter, we address static user-centric resource management strategies in heterogeneous network typically composed of areas covered via 3G/4G cellular macro-cells and WiFi hotspots (see illustration in Figure 3.1). Individual users can take advantage of the coexistence of the different wireless technologies for enhancing their perceived quality of service. A single user can decide to use one link or multiple links simultaneously based on system parameters such as throughput and energy consumption. We present in Section 3.1 the static network selection approach considering user objectives, rules and priorities to take decisions with learning in order to select either the WiFi or cellular network interface. In Section 3.2 we present the static traffic splitting as a multi-objective optimization problem capturing the tradeoff between throughput and energy consumption to decide on the amount of data to be received simultaneously over both links.

3.1 Static WiFi/Cellular Network Selection for Data Download

To meet the huge traffic growth, heterogeneous networks composed of wireless local area networks and cellular networks are used to provide higher capacity and coverage. When the two networks are available, selecting the best network for downloading data with minimum device energy consumption and high quality of service becomes a challenging issue especially that mobile devices have limited energy capacity. In the literature, the authors did not consider all the parameters known by the user that can affect the network selection decision. In addition, some results were based on simulations or arbitrarily generated values for factors affecting network selection decisions such as signal quality, network load and

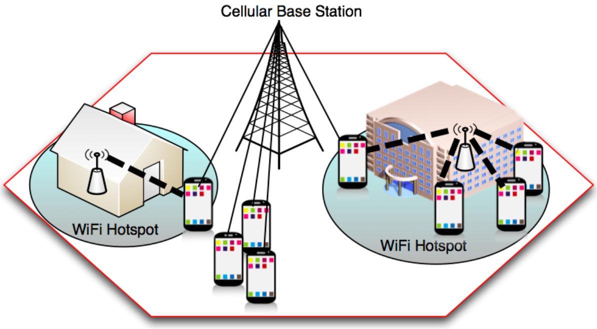


Figure 3.1: Heterogeneous network formed by cellular macro-cell and WiFi hotspots.

achievable data rate.

In our work, we present a learning-based approach for performing network selection based on real-network implementations. The main contributions are first, presenting an approach for building training data as a basis for machine learning of network selection and then developing the classification model for network selection. The model considers the features that affect the selection decision known by the user: availability of the networks, signal strength reflecting the channel quality, data size, battery life, speed of the user, location, and type of application. The training data set is based on experimental measurements of WiFi and 3G links using a Samsung Galaxy SII device. In particular, battery life, location and type of service determine the priority of the user in selecting the network that provides the user either the highest QoS, lowest energy consumption or highest energy efficiency based on pre-defined rules. To decide on the performance of each network in terms of rate and energy efficiency, the following attributes are used: user location, data size, and experimental values for WiFi and 3G signal strengths. For real-time network selection, the developed model uses decision tree classification. Testing the performance of the classifier using cross validation demonstrated high accuracy for selecting between WiFi and 3G networks.

This section is organized as follows. Section 3.1.1 describes the proposed network selection model including the proposed features for classification, the method for generating training model, and the machine learning model for network selection. The experimental results are described in Section 3.1.2. Network Selection limitations and challenges are presented in Section 3.1.3.

3.1.1 Learning-Based Approach for Network Selection

In this section, we present the details of the proposed learning-based approach for performing network selection. We present first the proposed features that can be available to every user's phone, and that can be used as discriminating attributes for deciding on the network. Section 3.1.1 covers the proposed approach for developing training data, which can then be used for deriving a classification model for network selection. The model for network selection is presented in Section 3.1.1.

Feature Selection

Selecting the network providing minimum device energy consumption and high quality of service to the user is a challenging issue due to the diversity of factors affecting the performance of the system. Some of these parameters cannot be determined by the user such as load on the networks, interference level, available resources and system capacity. Our proposed model considers all the parameters or attributes that affect the selection decision and that are well known by the user. These proposed features are:

- **Availability of the WiFi and 3G networks (A_{W} and A_{C}):** The two attributes A_{W} and A_{C} are defined to indicate the availability of the WiFi and 3G cellular networks, respectively. These attributes have binary possible values indicating whether the particular network is available or not.

- **Signal strength (S_{W} and S_{C}):** The two attributes S_{W} and S_{C} are real values in dBm and they represent the channel quality signal strength between the user and the WiFi access point and 3G cellular base station, respectively.

- **Data size (D):** This attribute is the size of the data in Bytes that the user requests to download.

- **Battery life (B):** The battery life attribute represents the remaining battery life percentage of the device. In this work, B is considered critical when the remaining battery life percentage is less than 20%.

- **Speed (V):** This attribute is a real value in m/s representing the speed of the user while downloading.

- **Location (L):** The location attribute is a binary attribute indicating whether the user is at home or outside home.

- **Type of application (T):** This attribute indicates if the type of the application is delay sensitive or not.

Table 3.1: Network Selection Rules

Features				**Decision**
V	**L**	**B**	**T**	**Rule**
$\geq$5m/s	-	-	-	3G
<5m/s	home	-	-	highest rate
<5m/s	not home	critical	not delay sensitive	lowest energy
<5m/s	not home	critical	delay sensitive	highest energy eff.
<5m/s	not home	not critical	-	highest energy eff.

Training Model Development

To develop training data that can be useful for the network classification model, supervised annotation of network class label is needed in association with every set of the features' values. We propose the rules shown in Table 3.1 for the annotation of network choice. Table 3.1 describes the rule choices under different scenarios when both WiFi and 3G networks are available. The first four columns of the table represent some of the features available to the mobile: the location L, the speed V of the user, the battery life B, and the service type T. The network selection rules are:

- If the user is moving faster than 5m/s, the user uses 3G cellular network since the WLAN coverage is limited and cannot manage high mobility [11].
- If the user is at home, battery life is neglected since the phone can be charged anytime. The link that offers the highest rate will be selected.
- If the user is not at home and the battery life is less than 20%, it is considered to be critical thus the network selected will be the network that offers lowest energy consumption if the service type is not delay sensitive.
- If the user is not at home and the battery life is critical, the network that offers highest energy efficiency is selected if the service is delay sensitive to provide a balance between rate and energy.
- If the battery life is not critical and the user is not at home, the user needs a good quality of service while increasing the battery life; therefore, the network with highest energy efficiency will be selected.

To derive annotation from the proposed rules, additional measures are needed to fire the different rules. These measures are: rate R, energy consumption E, and energy efficiency η. Figure 3.2 summarizes the proposed approach for building the training data set based on the selected features, rules and measurements.

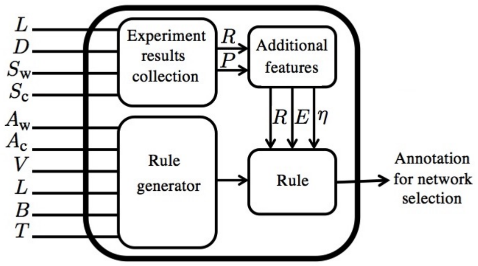


Figure 3.2: Basic components of building the training data for the proposed network selection model.

The figure shows the basic components of the network selection model. The following features: A_{W}, A_{C}, V, L, B and T are used for rule generation. While, the combination of other features: D, L, S_{W} and S_{C} are used to define the rate, energy, and energy efficiency accordingly. Combining the rules with the measurement results, the annotation for network selection is derived for each set of features' values, and ultimately leading to the complete training data.

The additional features are measured experimentally for generating training data only, and they are not available as features in actual classification. The measurements are collected under different scenarios of the following features: the location L, the signal strength S_{W}, S_{C} and the size D of the data that needs to be downloaded. The rate R and power consumed P are measured directly from the application and from a data acquisition device when experiments are conducted. Section 3.1.2 presents more details and sample results. Then, energy E and energy efficiency η are derived indirectly from measurements as follows:

$$E(\text{Joules}) = P(\text{W}) \cdot Time(\text{s}) = P(\text{W}) \cdot \frac{D(\text{bits})}{R(\text{bps})} \tag{3.1}$$

$$\eta(\text{bits/Joule}) = \frac{D(\text{bits})}{E(\text{Joules})} \tag{3.2}$$

To illustrate a sample of the data, Table 3.2 presents measurements that are collected experimentally under a specific set of conditions for L, S_{W}, S_{C} and D. In this scenario, the user is at home, close to the WiFi hotspot, having a bad 3G signal and needs to download a 1MB file. When the file is downloaded, the rate and power consumption are measured to determine the energy consumption and energy efficiency of each link.

In this scenario, the user has a speed lower than 5m/s and is at home. As a result, based on the rules in Table 3.1 and the collected measurements, the network with highest rate is selected. Combining the rule with the collected

Table 3.2: Features and Measurements: Sample Values

Features				Measurements R(Mbps) and P(W)			
L	D	$S_{\mathbf{W}}$	$S_{\mathbf{C}}$	$R_{\mathbf{W}}$	$P_{\mathbf{W}}$	$R_{\mathbf{C}}$	$P_{\mathbf{C}}$
home	1MB	-38dBm	-101dBm	1.8Mbps	1.044W	0.6Mbps	1.45W

measurements, the network that provides highest rate is WiFi. Thus, the annotation corresponding to this set of features is set to WiFi.

Network Selection Model

For the development of the network selection model, we propose supervised learning with the developed training data in Subsection 3.1.1. This is a standard step in machine learning, with several options for classification algorithms such as the use of decision trees, Naive Bayes, and Support Vector Machine (SVM) [96]. Once the model is developed, a new set of features available to the user can be fed to the model, and the real time decision is obtained on the network choice for downloading data based on the user status and network conditions. In this work, we propose the use of decision trees since the model gives a set of rules that can be logically evaluated for their relevance. The performance accuracy and evaluation of the classifier can be tested using standard methods such as cross validation. Details of the derived decision tree with our experimental measurements are presented in Section 3.1.2.

3.1.2 Experimental Results and Analysis

In this section, we present first the details of the experimental setup and scenarios conducted to collect the needed measurements. In Section 3.1.2, the sample measurements results are first presented and analyzed to demonstrate the validity of the measurements collected. Second, the method for building the training data based on the rules and measurements is presented and illustrated by several scenarios. The results for the network selection classification model based on decision tree classification are presented in Section 3.1.2.

Experimental Setup

To collect the data needed to build the training data set, the following setup was used. First, an Android application was developed on a Samsung Galaxy SII device. The purpose of the application was to download different data sizes while varying the device location, thus, changing the signal strength of WiFi and 3G. To measure performance under different conditions, six different locations were

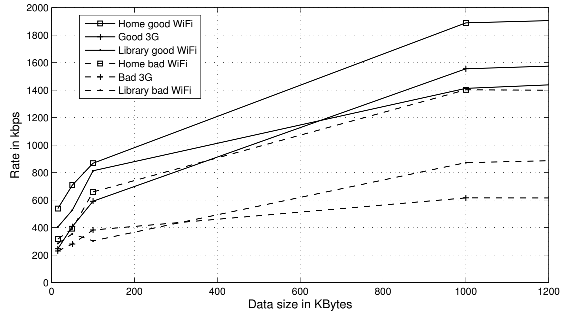


Figure 3.3: Data rate variation with 3G and WiFi signal strength and data size.

chosen: at home near the WiFi hotspot (home with good WiFi signal strength), at home far from the WiFi AP (home with bad WiFi signal strength), at the library where the network is loaded near the WiFi AP (library with good WiFi signal) and far from the WiFi AP (library with bad WiFi signal), indoor with bad 3G signal and outdoor with good 3G signal. In each scenario, the download data rate was obtained from the application while power consumption was measured using a data acquisition device (DAQ) monitored by LABVIEW. The energy consumption and energy efficiency were derived offline using (3.1) and (3.2), respectively.

Training Model Development Results

First, to show the validity of our experimental results, the measurements collected were analyzed. Then, the approach for building the training data set is presented. Figures 3.3, 3.4 and 3.5 capture the variation of rate, energy consumption and energy efficiency with respect to data size (Kbytes), location (home and library) and signal strength (near, far from the WiFi hotspot, indoor and outdoor). Figure 3.3 shows that rate increases when the user has better signal quality when closer to WiFi AP. These results were as expected since the transmission quality is affected by the channel between the transmitter and the receiver. The data rate increased with data size since the calculations took into consideration the connection setup time; however, data rate saturates when data size is larger than 1 MB in our measured scenarios. In addition, the data rate depends on the load on the WiFi network. In our model, we assumed that when the user is outside home, the WiFi network is considered to be loaded such as in a library environment. As expected, the WiFi rate showed lower values in a loaded environment. The results for power consumption showed that the receiving power of a Samsung Galaxy SII device was on average 1.044 Watts for WiFi and 1.45 Watts for 3G.

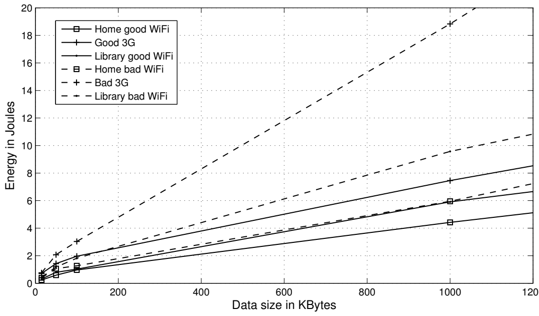


Figure 3.4: Energy consumption variation with 3G and WiFi signal strength and data size.

These values represent the power consumed by the mobile when the application is open and receiving and the brightness of the mobile is medium. The mobile power consumption is approximately constant since when receiving the mobile will be processing the data received. In general, the power consumed when 3G is used is more than the power consumed while receiving via WiFi due due to the additional processing requirements at the device level; these results are similar to the results provided in [97]. The energy consumption is then computed using (3.1). As shown in Figure 3.4, the energy consumption increased when the data size increased since it needs more time to download the data. The results also showed lower energy consumption when having good signal strength of WiFi and 3G since the rate is higher and, thus, the time needed to download data will be lower. Figure 3.5 shows the energy efficiency variations with respect to location and signal strength. The energy efficiency is computed using (3.2) and is a balance between data rate and power consumption. In our measurements, downloading data from WiFi with good signal strength was always more energy efficient than downloading data via 3G. The network selection depends on the rules discussed in Table 3.1. These rules needed additional measurements on rate, energy consumption and energy efficiency. The measurements presented above were used to get these additional information. To illustrate the method used for developing the training data set based on the rules and measurements, real examples are presented as follows. We assumed first that a user needs to download a 1 MB file from home having a speed lower than 5m/s. Based on the rules, if the user is at home, the link that provides the best rate is selected. Based on measurements, the rates provided by WiFi and 3G are compared. Table 3.3 presents sample measured data rate, energy consumption and energy efficiency for different combinations of L, S_{W}, S_{C} and D. Assuming the user is close to the WiFi AP, and have a bad 3G signal, the rate of WiFi is greater than rate of 3G

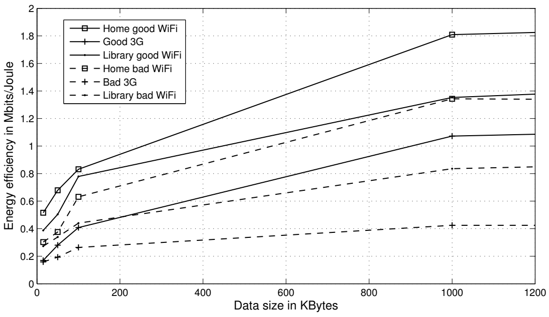


Figure 3.5: Energy efficiency variation with 3G and WiFi signal strength and data size.

therefore the annotation is WiFi. Considering another scenario where the user is outside home having a good 3G signal and connected to a far loaded WiFi AP. Assuming the user has a critical battery life and the application is not delay sensitive and following the rules, the least energy consuming link is selected which is in this case the 3G network as provided in Table 3.3.

Table 3.3: Data rate, energy consumption and energy efficiency when downloading a 1 MB data file from different locations with different signal strengths

L	S	R[Mbps]	E[J]	η[Mbits/Joule]
Home	good Wifi -39dBm	1.88	4.44	1.80
Home	bad WiFi -80dBm	1.40	5.96	1.34
Library	good WiFi -39dBm	1.41	5.92	1.35
Library	bad WiFi -80dBm	0.87	9.60	0.83
Outdoor	good 3G -61dBm	1.55	7.48	1.07
Indoor	bad 3G -101dBm	0.61	19.01	0.42

Based on the rules and measurements previously presented, the training data set is developed. It is composed of the following nine attributes: A_{W}, A_{C}, B, V, T, L, D, S_{W} and S_{C}. The training data set was formed by 7920 tuples representing the number of scenarios considered for different combinations of the attributes.

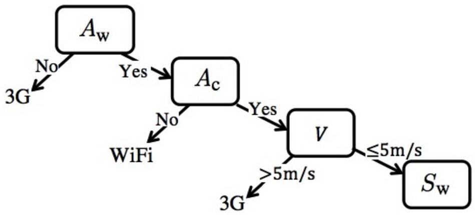


Figure 3.6: First four levels of the decision tree.

Network Selection Classification Results

The supervised training data set was imported into the RapidMiner data mining tool to generate the decision tree based on gain ratio [98]. As shown in Figure 3.6, the root attribute was A_W and then A_C. The tree showed that if WiFi is not available, the model selects 3G network. As expected from the rules used in the training data, the third highest level attribute is the speed of the user. If the two networks are available and the user's speed is more than 5m/s, 3G is selected. Based on the remaining attributes, the model makes the network selection decision by following the decision tree. The results of the decision tree were consistent with our observations and experimental results analysis. For instance, one path of the decision tree showed that WiFi was always selected when the 3G signal is less than -96dBm and WiFi signal is less than -70dBm. This classification is consistent with the measurements results that showed that in the case of having bad WiFi and 3G signals, WiFi provided better gains in term of rate, energy consumption and efficiency. The performance evaluation of the classifier was tested using cross validation, with 66% of the tuples in the training data set used as training data to build the decision tree and the remaining tuples used to test the decision tree. The classifier for the considered scenarios led to 99.77% average accuracy.

3.1.3 Static Network Selection Limitations and Challenges

To this extend, we considered a simplified version of network selection where transmission rates are estimated based on previous study on signal strength channel quality. Selecting the best network for downloading data provides moderate performance gains in terms of throughput and energy consumption, however, some applications can compensate some energy to gain high achievable rates when using traffic splitting over both link simultaneously. The features introduced in this work will be used for later consideration of dynamic and real-time network selection implementations. In addition, the main concern and challenge is to consider traffic splitting where the user can take advantage of simultaneous transmission over different links to achieve performance gains.

3.2 Static WiFi/Cellular Traffic Offloading for Data Download

In this section, we address user-centric traffic splitting in HetNets where a mobile device can simultaneously utilize both wireless interfaces to achieve performance gains. We propose a multi-objective traffic splitting approach that can be tailored to optimize different performance metrics and to capture existing energy-throughput tradeoffs. The amount of data sent over each interface is determined based on parameters that include both the effective data rate and the energy consumption over each link. Energy-throughput tradeoff results are presented including energy consumption minimization and throughput maximization results. The main contribution of this work is utilizing the energy-throughput tradeoff framework to develop optimized strategies that can guarantee a balance between energy consumption and throughput based on the user's needs in terms of application service requirements and mobile device battery life. Moreover, experimental measurements are used to determine values for key parameters in order to evaluate the proposed traffic splitting approach under realistic network conditions. This work has been published in proceedings of the IEEE Wireless Communications and Networking Conference (WCNC'14)[99].

This section is organized as follows. The system parameters are presented in Section 3.2.1. The proposed traffic splitting approach is detailed in Section 3.2.2. Experimental and performance results are presented and explained in Section 3.2.3. Static traffic splitting limitations and challenges are presented in Section 3.2.4.

3.2.1 System Parameters

When a mobile device has both cellular and WiFi connections available, the best link for data transmission or the best split ratio of data over the two links should be determined to optimize performance. The decision can be made at the user or network end. Each link provides the user with a specific data rate depending on the user's location, channel conditions, and access network load. Moreover, the energy consumed from the user's mobile device depends on the data rate, download time, data size, and wireless interface characteristics. In general, the main system parameters are:

- **Channel quality signal strength (S_{W} and S_{C}):** S_{W} and S_{C} represent the channel quality signal strength between the user and the WiFi access point and cellular base station, respectively.

- **Data size (D_{W}, D_{C} and D_{HN}):** D_{HN} is the total data size that the user requests to download. It will be split into the two links, thus, the data sent over WiFi and cellular links is D_{W} and D_{C}, respectively.
- **Time needed to download the data (T_{W} and T_{C}):** The time needed to receive D_{W} and D_{C} over WiFi and cellular links is T_{W} and T_{C} seconds, respectively.
- **Data rate (R_{W} and R_{C}):** The data rates over WiFi and cellular links are R_{W} and R_{C}, respectively, expressed in bits/second.
- **Power consumption (P_{W} and P_{C}):** The power consumed by the mobile device while receiving via WiFi and cellular links is P_{W} and P_{C} in Watts, respectively.
- **Energy consumption (E_{W} and E_{C}):** The energy consumed when the mobile device is using WiFi and cellular links is E_{W} and E_{C} Joules, over download duration T_{W} and T_{C}, respectively; it can be computed as follows: $E = PT$.

The heterogeneous network characteristics with traffic splitting can be defined as follows:

- The total amount of data downloaded D_{HN} is equal to the summation of data sent via both links as follows:

$$D_{\mathrm{HN}} = D_{\mathrm{W}} + D_{\mathrm{C}} \tag{3.3}$$

- The HetNet total download time using traffic splitting is equal to the maximum download time between both links:

$$T_{\mathrm{HN}} = \max(T_{\mathrm{W}}, T_{\mathrm{C}}) \tag{3.4}$$

- The total HetNet rate R_{HN} with traffic splitting over both links is given by:

$$R_{\mathrm{HN}} = \frac{D_{\mathrm{HN}}}{T_{\mathrm{HN}}} = \frac{D_{\mathrm{HN}}}{\max(T_{\mathrm{W}},\, T_{\mathrm{C}})} = \frac{D_{\mathrm{HN}}}{\max\left(\frac{D_{\mathrm{W}}}{R_{\mathrm{W}}}, \frac{D_{\mathrm{C}}}{R_{\mathrm{C}}}\right)} \tag{3.5}$$

- The energy consumed by the device when using traffic splitting is assumed to be equal to the summation of the energy consumed by each link as follows:

$$\begin{aligned} E_{\mathrm{HN}} &= E_{\mathrm{W}} + E_{\mathrm{C}} = P_{\mathrm{W}}T_{\mathrm{W}} + P_{\mathrm{C}}T_{\mathrm{C}} \\ &= \frac{P_{\mathrm{W}}}{R_{\mathrm{W}}}D_{\mathrm{W}} + \frac{P_{\mathrm{C}}}{R_{\mathrm{C}}}D_{\mathrm{C}} \end{aligned} \tag{3.6}$$

3.2.2 Multi-Objective Optimization Approach for Traffic Splitting

We present a multi-objective approach for traffic splitting in cellular/WiFi heterogeneous networks that can be tailored to throughput and energy performance metrics, and can be used to capture existing energy-throughput tradeoffs. Using the link characteristics presented in Section 3.2.1, the data splitting ratio between both interfaces can be determined by solving the energy-throughput optimization problem as follows:

$$\begin{aligned} \underset{D_\mathrm{W}, D_\mathrm{C}}{\text{minimize}} \quad & (1-\alpha)E_\mathrm{HN} - \alpha R_\mathrm{HN} \\ \text{subject to} \quad & D_\mathrm{W} \geq 0, \quad D_\mathrm{C} \geq 0 \\ & D_\mathrm{HN} = D_\mathrm{W} + D_\mathrm{C} \end{aligned} \tag{3.7}$$

The problem is a bi-objective optimization problem for traffic splitting aiming to maximize the HetNet throughput R_HN while minimizing the energy consumption E_HN of the device. The throughput R_HN and energy consumption E_HN are expressed in (3.5) and (3.6), respectively. The decision variables to be determined are D_W and D_C, the amount of data to be sent over WiFi and cellular links, respectively. The problem is subjected to the following constraints: i) D_W and D_C should be positive, and ii) the total amount of data downloaded D_HN should be equal to the summation of data sent via both links. The α parameter gives bias weights to R_HN and E_HN in the objective function; it varies between 0 and 1.

The optimization problem presented in (3.7) varies with α as follows: i) when α is 1, the objective will be maximizing R_HN presented in Section 3.2.2, ii) when α is 0, the objective will be minimizing E_HN described in Section 3.2.2, and iii) when α varies between 0 and 1, energy-throughput tradeoff solutions are presented in Section 3.2.2.

User Throughput Maximization

Several real-time applications require high bit rates to give the user the best quality of experience. Therefore, the first approach is to maximize the HetNet throughput R_HN in (3.5) under the given constraints. The problem in (3.7) will be then reduced to maximize R_HN when α is equal to 1; this is equivalent to minimizing the denominator max $\left(\frac{D_\mathrm{W}}{R_\mathrm{W}}, \frac{D_\mathrm{C}}{R_\mathrm{C}}\right)$ in (3.5), since the total data size D_HN is fixed. The problem can be solved by replacing the denominator by a new variable g and additional constraints as follows:

$$\begin{aligned} \underset{D_{\mathrm{W}}, D_{\mathrm{C}}}{\text{minimize}} \quad & g \\ \text{subject to} \quad & D_{\mathrm{W}} \geq 0, D_{\mathrm{C}} \geq 0 \\ & D_{\mathrm{HN}} = D_{\mathrm{W}} + D_{\mathrm{C}} \\ & \frac{D_{\mathrm{W}}}{R_{\mathrm{W}}} \leq g \text{ and } \frac{D_{\mathrm{C}}}{R_{\mathrm{C}}} \leq g \end{aligned} \tag{3.8}$$

The best solution is obtained by minimizing the total transmission time T_{HN} which is equal to $\max(T_{\mathrm{W}},\, T_{\mathrm{C}})$. Therefore, the solution obtained is to send data over equal time durations on both links and, thus, T_{W} will be equal to T_{C}. The ratio Γ between D_{W} and D_{C} will be as follows:

$$\Gamma = \frac{D_{\mathrm{W}}}{D_{\mathrm{C}}} = \frac{R_{\mathrm{W}}}{R_{\mathrm{C}}} \tag{3.9}$$

Accordingly, the amount of data sent over WiFi to achieve maximum HetNet throughput denoted by $D_{\text{W-maxR}}$ is proportional to the data rate as follows:

$$D_{\text{W-maxR}} = \frac{\Gamma}{1+\Gamma} D_{\mathrm{HN}} = \frac{R_{\mathrm{W}}}{R_{\mathrm{W}} + R_{\mathrm{C}}} D_{\mathrm{HN}} \tag{3.10}$$

Energy Consumption Minimization

When α is equal to 0, the problem in (3.7) becomes minimizing the energy consumed from the battery of the mobile device expressed in (3.6). Minimum energy consumption is achieved when downloading all the data over the more energy efficient link without traffic splitting. Maximizing the energy efficiency, which is the ratio of the rate over the energy consumed, is solved by minimizing the energy consumption since the total data size to be downloaded is fixed; the link with lowest energy consumption will be used for data transmission.

Joint Energy-Throughput Considerations

In this section, we extend the traffic splitting approach to optimize an objective function that jointly captures the HetNet throughput and energy consumption as presented in (3.7). Maximizing R_{HN} can be solved by minimizing its denominator g, as presented in Section 3.2.2. The objective function will be:

$$\underset{D_{\mathrm{W}}, D_{\mathrm{C}}}{\text{minimize}} \quad (1-\alpha)\left(\frac{P_{\mathrm{W}}}{R_{\mathrm{W}}} D_{\mathrm{W}} + \frac{P_{\mathrm{C}}}{R_{\mathrm{C}}} D_{\mathrm{C}}\right) + \alpha g \tag{3.11}$$

It can be shown that the solution to this optimization problem varies between two cases only: either sending all data over the most energy efficient link, when α

is below a threshold Θ, to minimize the energy consumption; or sending data with equal transmission time between both links which is the solution of maximizing the throughput, when α is greater than Θ. Developing the objective function in (3.11) and combining it with (3.3) and (3.4), Θ can be derived as follows:

$$\Theta = \begin{cases} \dfrac{\frac{P_{\text{C}}}{R_{\text{C}}} - \frac{P_{\text{W}}}{R_{\text{W}}}}{\frac{P_{\text{C}}}{R_{\text{C}}} - \frac{P_{\text{W}}}{R_{\text{W}}} + \frac{1}{R_{\text{W}}}} & \frac{R_{\text{W}}}{P_{\text{W}}} \geq \frac{R_{\text{C}}}{P_{\text{C}}} \\ \\ \dfrac{\frac{P_{\text{W}}}{R_{\text{W}}} - \frac{P_{\text{C}}}{R_{\text{C}}}}{\frac{P_{\text{W}}}{R_{\text{W}}} - \frac{P_{\text{C}}}{R_{\text{C}}} + \frac{1}{R_{\text{C}}}} & \text{otherwise} \end{cases} \tag{3.12}$$

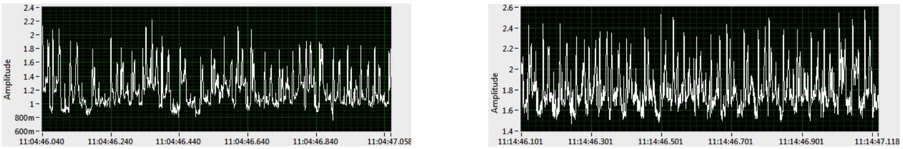
Since the optimal solution tends towards favoring one of two cases, a near-optimal solution is needed to guarantee a balance between energy consumption and throughput based on the user's needs in terms of application service requirements and mobile device battery life. For this reason, the energy-throughput tradeoff in HetNets with traffic splitting is quantified by expressing the HetNet energy in terms of the HetNet throughput. The HetNet throughput in (3.5) can be expressed in terms of D_{W} as follows:

$$R_{\text{HN}} = \begin{cases} \dfrac{R_{\text{C}} D_{\text{HN}}}{D_{\text{HN}} - D_{\text{W}}} & 0 \leq D_{\text{W}} \leq D_{\text{W-maxR}} \\ \\ \dfrac{R_{\text{W}} D_{\text{HN}}}{D_{\text{W}}} & D_{\text{W-maxR}} \leq D_{\text{W}} \leq D_{\text{HN}} \end{cases} \tag{3.13}$$

$D_{\text{W-maxR}}$ is the data size downloaded over WiFi that leads to maximum HetNet throughput, which occurs when the two links are used for the same duration. If D_{W} is less than $D_{\text{W-maxR}}$, more data is sent over the cellular link; thus, the denominator of (3.5) will be equal to the ratio $D_{\text{C}}/R_{\text{C}}$. If D_{W} is greater than $D_{\text{W-maxR}}$, the WiFi link is used for longer duration; therefore, the denominator will be equal to $D_{\text{W}}/R_{\text{W}}$. Combining (3.3), (3.6) and (3.13), the HetNet energy consumption of the device can be represented in terms of R_{HN} as follows:

$$E_{\text{HN}} = \begin{cases} \dfrac{-a}{R_{\text{HN}}} + b & 0 \leq D_{\text{W}} \leq D_{\text{W-maxR}} \\ \\ \dfrac{c}{R_{\text{HN}}} + d & D_{\text{W-maxR}} \leq D_{\text{W}} \leq D_{\text{HN}} \end{cases} \tag{3.14}$$

where a, b, c and d are parameters that depend on the WiFi and cellular link



(a) Power consumed (amplitude) by the device when using WiFi.

(b) Power consumed (amplitude) by the device when using 3G.

Figure 3.7: Power consumption by Samsung SIII device while receiving data over WiFi and 3G links.

characteristics:

$$a = \left(\frac{P_{\mathrm{W}}}{R_{\mathrm{W}}} - \frac{P_{\mathrm{C}}}{R_{\mathrm{C}}}\right) R_{\mathrm{C}} D_{\mathrm{HN}} \tag{3.15}$$

$$b = \left(\frac{P_{\mathrm{W}}}{R_{\mathrm{W}}} - \frac{P_{\mathrm{C}}}{R_{\mathrm{C}}}\right) D_{\mathrm{HN}} + \frac{P_{\mathrm{C}}\ D_{\mathrm{HN}}}{R_{\mathrm{C}}} \tag{3.16}$$

$$c = \left(\frac{P_{\mathrm{W}}}{R_{\mathrm{W}}} - \frac{P_{\mathrm{C}}}{R_{\mathrm{C}}}\right) D_{\mathrm{HN}} R_{\mathrm{W}} \tag{3.17}$$

$$d = \frac{P_{\mathrm{C}}\ D_{\mathrm{HN}}}{R_{\mathrm{C}}} \tag{3.18}$$

Using (3.14), the decision for traffic splitting can be tailored towards satisfying user's requirements in terms of throughput and energy consumption and providing near-optimal traffic splitting decisions based on the mobile battery status and type of application. More details are presented in Section 3.2.3.

3.2.3 Results and Analysis

To validate the presented multi-objective traffic splitting approach under realistic conditions, experimental measurements are used to determine WiFi and cellular key link parameters, such as effective download rate and energy consumed per second. The obtained link parameter values are then used to quantify and analyze the energy-throughput tradeoffs of HetNet traffic splitting in various scenarios.

Parameter Setting Using Experimental Measurements

Experimental measurements were conducted to capture the effect of signal strength and traffic load on the effective download rate and energy consumption. An Android application was developed on the Samsung Galaxy SIII device to download data from an HTTP server via WiFi (802.11b) and 3G cellular links at different locations. In each scenario, the data rate was obtained from the application while power consumption was measured using a data acquisition device from National

Instruments monitored via a LABVIEW application. Figures 3.7(a) and 3.7(b) show the plots for the power consumed by the mobile device when using WiFi and 3G, respectively. These plots demonstrate that the mobile device consumes more energy when using 3G interface than when using the WiFi interface. The average power consumed is 1.307 Watts for WiFi and 1.859 Watts for 3G.

The effective download data rate is affected by the link's traffic load and signal strength (SS), which depends on the user's location with respect to the WiFi AP or 3G BS. The data rate was measured in six locations: i) user is at home near WiFi AP (home with good WiFi SS of -39 dBm) and far from AP (home with bad WiFi SS of -80 dBm); ii) user is at library where the WiFi AP is loaded near (library with good WiFi SS of -40 dBm) and far (library with bad WiFi SS of -80 dBm); iii) indoor with bad 3G SS of -101 dBm and outdoor with good 3G SS of -61 dBm. The different data rates achieved when downloading a 6 MB file from different locations are summarized in Table 3.4. For these case studies, the results show that higher rates are achieved when downloading over WiFi with close proximity to the AP, especially in a home scenario with low access network load. Moreover, they demonstrate the notable variation in download bit rate between WiFi and 3G for different signal strength levels.

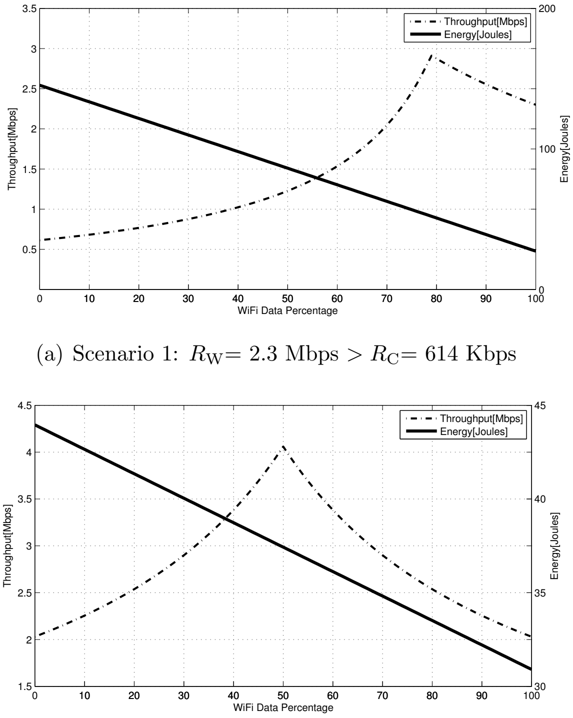
Table 3.4: Data rates when downloading a 6 MB data file over WiFi and 3G in different locations

Location	**Data Rate**
WiFi home (good WiFi signal strength -39 dBm)	2.299 Mbps
WiFi home (bad WiFi signal strength -80 dBm)	1.317 Mbps
WiFi library (good WiFi signal strength -40 dBm)	2.031 Mbps
WiFi library (bad WiFi signal strength -80 dBm)	1.221 Mbps
3G outdoor (good 3G signal strength -61 dBm)	2.030 Mbps
3G indoor (bad 3G signal strength -101 dBm)	0.614 Mbps

Performance Results: Maximum HetNet Throughput and Minimum HetNet Energy Consumption

This section presents analysis for two HetNet scenarios for downloading a 6 MB data file showing traffic splitting for maximum HetNet throughput and minimum energy consumption. In the first scenario, the user has higher data rate when connected to home WiFi AP. In the second scenario, the user has equal WiFi and 3G rates.

Scenario 1: The user is at home near the WiFi AP, and connected to 3G with a low signal strength. The WiFi data rate is 2.3 Mbps and is larger than the 3G rate which is 614 Kbps. Figure 3.8(a) plots HetNet throughput and energy



(a) Scenario 1: R_{W}= 2.3 Mbps > R_{C}= 614 Kbps

(b) Scenario 2: R_{W}= R_{C}= 2.03 Mbps

Figure 3.8: The left and right y-axes measure HetNet throughput in Mbps and HetNet energy consumption in Joules, respectively, with respect to the percentage of data sent over WiFi when downloading a 6 MB file with P_{W}= 1.307 Watts and P_{C}= 1.859 Watts.

consumption as a function of the HetNet traffic splitting ratio represented by the percentage of data sent over WiFi. The left and right y-axes measure the HetNet throughput in Mbps and the energy consumption in Joules, respectively. When the percentage of data sent over WiFi increases, the throughput increases while the energy consumption decreases which enhances the quality of service at the user end. The maximum achieved HetNet throughput is 2.91 Mbps when the percentage of data WiFi is 78.93% which can be obtained by the solution of the throughput maximization objective discussed in Section 3.2.2. Using WiFi alone is more energy efficient but provides less throughput compared to the case when using WiFi and 3G simultaneously. Using both links together was able to maximize the throughput with a tradeoff cost in terms of energy consumption.

Scenario 2: Assuming loaded WiFi and good 3G signal strength scenarios, the WiFi and 3G data rates are equal to 2.03 Mbps. As shown in Figure 3.8(b), the data rate doubles to be 4.06 Mbps when data is split equally over the two links, while the energy consumption is higher than the energy consumed when WiFi is used only.

Performance Results: Energy-Throughput Tradeoffs and Traffic Splitting Strategies

Figure 3.9 plots HetNet energy consumption versus HetNet throughput based on (3.14) by varying D_{W} from 0 to the total data size D_{HN}. The rates R_{W} and R_{C} are assumed to be equal to 2 Mbps, P_{W}= 1.307 Watts, P_{C}= 1.859 Watts and D_{HN}= 6 MB. The least energy consuming download is at point $E_{\min}$ when all data is sent over WiFi which is in this case the more energy efficient link. When D_{W} increases, the energy consumption and throughput increase to reach maximum rate at 4 Mbps where the data splits equally between the two links. Downloading more traffic over the cellular link leads to higher energy consumption while the same HetNet throughput can be achieved by another splitting ratio. For instance, the highlighted points P_1 and P_2 provide the same HetNet throughput while P_1 is more energy consuming.

The energy-throughput tradeoff aims to maintain a high throughput while keeping the energy consumption low. Therefore, the lower curve (solid line) in Figure 3.9 represents the more energy efficient splitting ratios ranging between minimum HetNet energy consumption at point $E_{\min}$, and the energy consumed when data is split equally between both links providing maximum HetNet throughput at $R_{\max}$.

When the channel conditions vary, the energy consumption and throughput will change accordingly. Three scenarios are considered in Figure 3.10 to download a 6 MB data file while R_{C} is assumed to be fixed equal to 2 Mbps. In the first plot, where R_{W}=R_{C}, the plot is symmetrical since maximum rate is achieved when data is split equally between the two links. When R_{W} increases, the maximum throughput increases while the minimum energy consumed decreases.

The highlighted points U_1, U_2 and U_3 are the ideal operational points for the each of three plots, respectively. The ideal operational point is the point where energy consumption is minimum and throughput is maximum. The ideal operational point is shown to be always closer to the maximum throughput solution than to the minimum energy consumption solution. From a practical implementation perspective, the selection of the most suitable operational point depends on the applications type and quality requirements in addition to the available battery capacity in the mobile device. On one hand, several applications require a guaranteed bit rate. Therefore, the strategy for traffic splitting needs to provide the minimum throughput required even if the device's battery life is critical, while it can offer higher rates when the battery capacity is high.

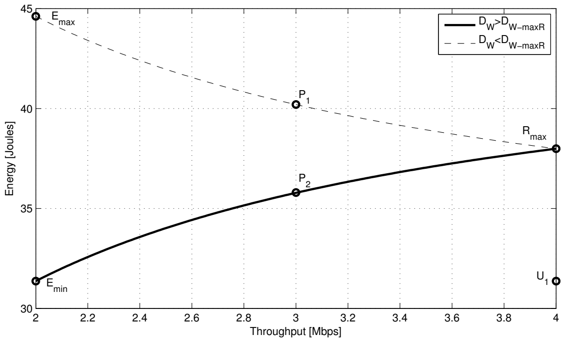


Figure 3.9: The HetNet energy consumption variation with respect to HetNet throughput with D_{HN}= 6 MB, R_{W}= R_{C}= 2 Mbps.

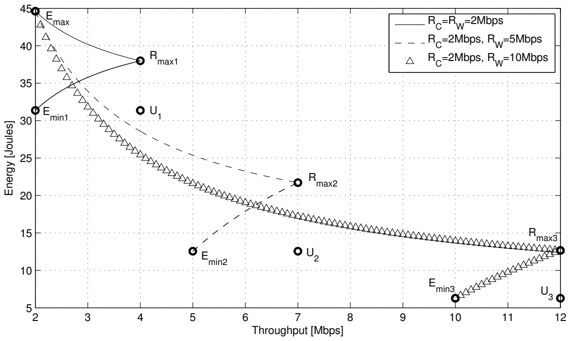


Figure 3.10: The HetNet energy consumption variation with respect to HetNet throughput when downloading a 6 MB file.

On the other hand, the most suitable traffic splitting strategy can be determined based on the available battery capacity. For instance, a target battery life can determine the maximum possible HetNet throughput. The target battery life can be represented by the number of remaining hours of device operation assuming the mobile device keeps on receiving continuously. The remaining hours of operation can be calculated as the ratio of the available battery capacity (in mWh) and the average power consumed (in Watts). The average power consumed can be computed by the ratio of HetNet energy and the total transmission time. Therefore, the remaining hours of operation depends on the available battery capacity and on the traffic splitting ratio which determines the HetNet through-

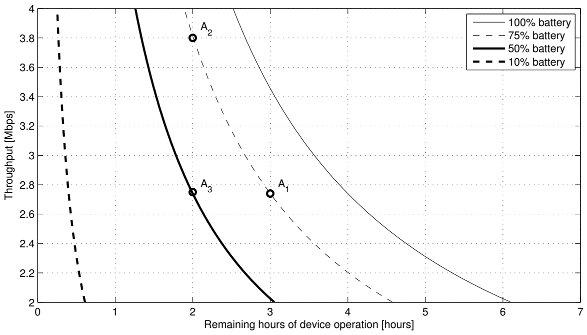


Figure 3.11: The HetNet throughput variation with respect to the remaining hours of device operation, with D_{HN}= 6 MB , R_{C}= R_{W}= 2 Mbps.

put and energy consumption. Figure 3.11 shows the HetNet throughput versus the remaining hours of device operation for different available battery capacity levels. The device consumes more power to provide higher HetNet throughput which decreases the battery life and reduces the remaining hours of device operation. When the battery is fully charged (100%), the Samsung Galaxy SIII device offers a capacity of 2100 mAh using a voltage of 3.8 V, which corresponds to an available battery capacity of 7980 mWh.

The most suitable traffic splitting strategy can be determined based on a target remaining hours of device operation at a specific available battery capacity. For instance, if the mobile device battery is 75% charged, and the user needs its battery to last longer than three hours, based on the highlighted point A_1 in Figure 3.11, the traffic splitting will be performed to obtain a maximum HetNet throughput of 2.74 Mbps. Therefore, based on (3.5), the 6 MB data file is split as follows: 35 Mbits and 13 Mbits will be sent over WiFi and 3G cellular links, respectively. If the target battery life duration is two hours, higher HetNet throughput can be achieved. The highlighted point A_2 shows 3.8 Mbps HetNet throughput. If the battery capacity is 50% charged, the maximum allowed HetNet throughput to maintain two hours of battery life with continuous download, is 2.75 Mbps as indicated by the highlighted point A_3. When the battery life is critical (10% charged), lower rates are recommended to extend the device's battery life.

3.2.4 Static Traffic Offloading Limitations and Challenges

In this work, we addressed the operation of cellular/WiFi heterogeneous networks with network selection and traffic splitting. The model presented is static with no parameter variation with respect to time. The decision is made only once at the beginning of the data download. However, in real-time applications such as video streaming, the data is dynamically downloaded, buffered and played. Therefore, we present in Chapter 4 real-time optimization formulations considering dynamic variation of transmission parameters such as data rates, fast solutions and decision making for data splitting over two interfaces to provide the user with better quality of experience.

Chapter 4

Dynamic Cellular/WiFi Traffic Splitting for Video Streaming with QoE Considerations

As presented in Chapter 3, traffic splitting in HetNets was able to provide higher throughput with a tradeoff cost in terms of energy consumption. This demonstrates the potential gains of traffic splitting in HetNets and shows the need for intelligent dynamic strategies in practical scenarios, that are especially customized to real-time streaming applications.

In this chapter, we address dynamic cellular/WiFi traffic splitting for video streaming with quality of experience considerations. Due to the real-time and fast aspects of video streaming applications, higher quality of experience and satisfaction are required. For these reasons, higher bandwidth needs to be allocated to the user by allowing the use of multiple wireless interfaces simultaneously. This will increase the energy consumption and cost for data plans. Previous work in this field has considered improving throughput and reducing energy consumption, but did not consider simultaneously quality of experience as perceived by the end user. In our work, we focus on dynamic traffic splitting decisions in cellular/WiFi HetNets to provide the user with a balance between high quality of experience, low energy consumption and delay while using video on demand streaming applications. The main contributions are first, developing an optimized multi-objective traffic splitting solution as a function of the dynamic variation of various system parameters. In contrast to the literature, our proposed approach does not focus only on throughput and energy consumption, but also considers user quality of experience based on ITU-T P.1201 standard to better capture the video quality as perceived by the end user. The proposed approach will determine in real-time how to intelligently split traffic between both wireless interfaces and decide dynamically on the amount of data to be sent over each wireless link. Real-time traffic splitting decisions are performed based on Lyapunov drift-plus-penalty optimization utility functions aiming at achieving high quality end-user experience and minimizing energy consumption while stabilizing network queues. The proposed dynamic traffic splitting with delay-power-QoE (TS-PQ) balance approach is user-centric and runs in the background at the user side without any intervention from the network or the server, and without performing any changes to the cellular/WiFi standards. The performance is evaluated using simulations based on parameters determined via experimental measurements on mobile devices for video streaming and using our own test bed implementation under realistic conditions. Results show the performance effectiveness of the proposed traffic splitting approach in terms of throughput, delay, queue stability, energy consumption and quality of user experience by monitoring the frequency and lengths of video stalls.

This chapter is organized as follows. The potential gains of traffic splitting are presented in Section 4.1. The system model is presented in Section 4.2. The proposed traffic splitting approach is detailed in Section 4.3. Performance results are presented and explained in Section 4.4. Test bed implementation and results are presented in Section 4.5. Finally, dynamic traffic splitting limitations and challenges are drawn in Section 4.6.

4.1 Motivating the Benefits of Traffic Splitting in HetNets

To demonstrate the potential gains of traffic splitting in heterogeneous networks, a realistic toy example for video on demand transmission is presented in Figure 4.1. Simulations are conducted using MATLAB to stream a video using different strategies: (1) WiFi only (WO), (2) cellular only (CO), and (3) using both links simultaneously (TS-S), and their performance in terms of average throughput, total energy consumption, frequency of stalls and length, and satisfaction metric evaluated based on ITU-T P.1201 (2013) QoE metric (4.15) (details are presented in Section 4.3.2). Figure 4.1 shows the three different data transmission decision strategies for three consecutive 117 KB data download time slots. The video has a size of 7 MBytes, duration of 60 seconds, and frame rate of 25 fps. The arrival rate will be 117 KBytes every second. At each time slot of duration 1 second, the mobile device will make decision on the links to use for downloading data based on the selected strategy. if the download rate is less than the video arrival rate, the user will experience buffering events. The video data is not lost and the frames are not skipped. Instead, they are delayed when stalls happen.

In the top-most plot, the data is always sent over WiFi. The average throughput over WiFi was higher than the average throughput provided by cellular link while the total energy consumed is the lowest comparing to the other two scenarios where data is sent over cellular only or split between the two links.

In the lowest plot, splitting the traffic by using both links simultaneously

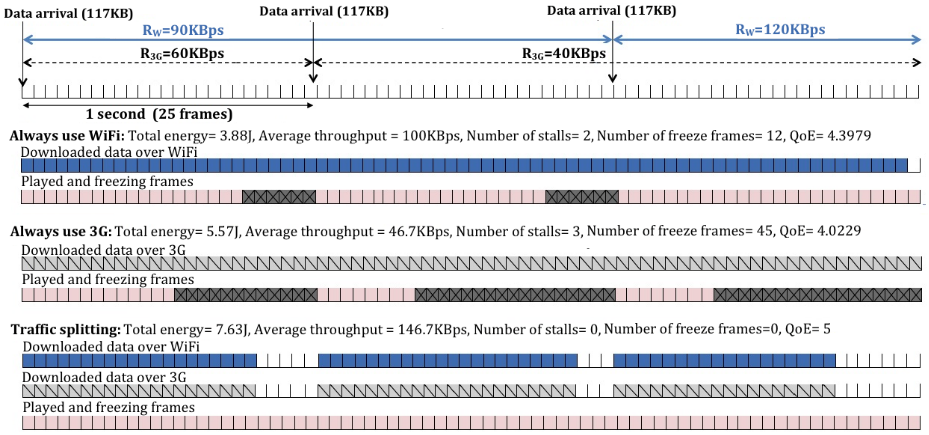

Data arrival (117KB)
Data arrival (117KB)
Data arrival (117KB)
R_W=90KBps
R_{3G}=60KBps
R_{3G}=40KBps
R_W=120KBps
1 second (25 frames)
Always use WiFi: Total energy= 3.88J, Average throughput = 100KBps, Number of stalls= 2, Number of freeze frames= 12, QoE= 4.3979
Downloaded data over WiFi
Played and freezing frames
Always use 3G: Total energy= 5.57J, Average throughput = 46.7KBps, Number of stalls= 3, Number of freeze frames= 45, QoE= 4.0229
Downloaded data over 3G
Played and freezing frames
Traffic splitting: Total energy= 7.63J, Average throughput = 146.7KBps, Number of stalls= 0, Number of freeze frames=0, QoE= 5
Downloaded data over WiFi
Downloaded data over 3G
Played and freezing frames

Figure 4.1: Toy example illustrating the benefits of traffic splitting by comparing three scenarios of: (i) always use WiFi, (ii) always use cellular, (iii) traffic splitting while using the two links simultaneously, for downloading three consecutive 117 KB data blocks. The Figures show the performance in terms of average throughput, total energy consumption, frequency of stalls and length (freeze frames are marked with X), and QoE metric evaluated based on (4.15).

provides higher throughput while consuming more energy compared to using WiFi and cellular alone. The results showed that the user experience stalls and freezing frames while watching over WiFi and cellular only, however, no stalls are experienced when traffic splitting is used. The estimated QoE is 5 when using traffic splitting while it is lower 4.3979 and 4.0229 when using WiFi only and cellular only, respectively.

This demonstrates the potential gains of traffic splitting in HetNets and proves its ability to provide high performance in terms of throughput and user satisfaction with a trade of in energy consumption. The results emphasize the need for an optimized decision making approach to achieve target performance tradeoff and motivate our work for proposing better solutions for designing an optimized approach to determine the best cellular/WiFi resource management strategy considering traffic splitting decisions in every time slot. Many questions arise for selecting the best suitable decision at each time slot: (1) What are the available interfaces and transmission strategies? (2) What are the effects of using each strategy on the device power consumption, queue length and user satisfaction? (3) What is the best strategy and the amount of data to be transmitted over each interface that reduces power consumption, queue backlog length, number and length of stalls, and maximizes user QoE? (4) Is it possible to provide a device centric approach performing autonomously without any intervention or change in the standards? (5) What will be the gains under practical implementation and operational conditions?

4.2 System Model and Parameters

Our proposed method decides on behalf of the user to use one link or multiple links simultaneously based on system parameters such as throughput, energy consumption, quality of experience and cellular/WiFi links characteristics. Therefore, deciding on the best download strategy needs to be dynamic to provide the user with the best performance at discrete time sample points, represented by time slots. The notion of time slots is introduced to handle discretization of the real-time aspect of the system and handle the queuing theory operation of the proposed approach. It provides practical feasibility to make decisions periodically every time slot based on data collected and previous actions.

Accordingly, at every time slot t, when a mobile device has both cellular and WiFi connections available, the best link for data download or the best split ratio of data over the two links should be determined to optimize performance and provide the best balance between QoE, energy consumption and delay.
In general, the main system parameters are:

- **Time slot duration (T_s)**: T_s is the time slot duration, in seconds, representing how often the decision is taken.

- **Resource management solution ($L[t]$):** $L[t]$ is the possible traffic splitting decision at time slot t, which can be one of the following: (1) WiFi only, (2) cellular only, (3) both links simultaneously, and (4) no transmission.

- **Strategy (ℓ):** The index ℓ represents one of the possible resource management strategies represented by $L[t]$. ℓ represents the index W when WiFi only is selected, C when cellular only is selected and WC when both links are used simultaneously. Based on the selected strategy ℓ, the amount of data to be sent over WiFi and cellular links, respectively during time slot duration T_s, can be determined.

- **Transmission data rate ($R_\ell[t]$):** $R_\ell[t]$ represents $R_\mathrm{W}[t]$ the estimated data rate over WiFi link only, $R_\mathrm{C}[t]$ over cellular link only, and $R_\mathrm{WC}[t]$ when using both links simultaneously at time slot t, expressed in bits/second. Note that $R_\mathrm{WC}[t] = R_\mathrm{W}[t] + R_\mathrm{C}[t]$.

- **Power consumption ($P_\ell[t]$):** $P_\ell[t]$ represents the estimated power consumed by the mobile device while receiving via WiFi, cellular or both links simultaneously, P_W, P_C and P_WC in Watts, respectively. Note that $P_\mathrm{WC}[t] = P_\mathrm{W}[t] + P_\mathrm{C}[t]$.

- **Arrival rate ($A[t]$):** $A[t]$ represents the amount of data, in bits, that arrives to the user's queue at the server from the application layer within time slot t.

- **Cost ($C_\ell[t]$):** $C_\ell[t]$ represents the estimated penalty function and cost in terms of power consumption and QoE degradation when choosing resource management strategy ℓ at time slot t. The selected resource management strategy will determine the cost of the decision at time slot t.

- **QoE metric ($\phi_\ell[t]$):** $\phi_\ell[t]$ represents the expected quality of experience and user satisfaction when choosing download strategy ℓ at time slot t. $\phi_\ell[t]$ represents $\phi_{\mathrm{W}}[t]$, $\phi_{\mathrm{C}}[t]$, $\phi_{\mathrm{WC}}[t]$ when WiFi only, cellular only, and simultaneous use of both links are selected, respectively. $\phi_\ell[t]$ is expressed as mean opinion score ranging from 1 to 5.

- **Number of stalls ($N_\ell[t]$):** $N_\ell[t]$ represents the predicted number of stalls or rebuferring events that the user is expected to experience if resource management strategy ℓ is used at time slot t. If the transmission rate is less than the required amount of data to be played, the user will experience a re-buffering event during time slot t; in this case, $N_\ell[t] = N[t-1]+1$.

- **Average stalls length ($L_\ell[t]$):** $L_\ell[t]$ represents the average length of stalls that the user is expected to experience if download strategy ℓ is used at time slot t. $L_\ell[t]$ considers the length of all the previous stalls experienced by the user in addition to the expected stall length at time slot t.

The performance parameters are:

- **Transmission data ($\mu_\ell[t]$):** $\mu_\ell[t]$ represents the amount of data that has been transmitted in time slot t over WiFi and cellular links $\mu_{\mathrm{W}}[t]$ and $\mu_{\mathrm{C}}[t]$, respectively, and on both links simultaneously $\mu_{\mathrm{WC}}[t]$, expressed in bits.

- **Transmission data ($\mu[t]$):** $\mu[t]$ represents the total amount of data that has been transmitted till time slot t, expressed in bits.

- **Queue backlog ($Q[t]$):** The queue backlog $Q[t]$ represents the amount, in bits, of unfinished work as data not being downloaded yet at the beginning of time slot t and can be expressed as follows:

$$Q[t+1] = Q[t] - \mu_\ell[t] + A[t] \tag{4.1}$$

- **Video data played ($Y[t]$):** $Y[t]$ represents the amount of video data played till time slot t, expressed in bits.

- **Video data downloaded but not yet played ($D[t]$):** $D[t]$ represents the amount of video data downloaded but not yet played till time slot t, expressed in bits. $D[t]$ can be computed as follows: $D[t] = \max(\mu[t] - Y[t], 0)$.

- **QoE ($\phi[t]$):** $\phi[t]$ represents the quality of experience metric reflecting the user satisfaction at time slot t. QoE metric does not capture subjective quality of experience, however, $\phi[t]$ will reflect the user satisfaction based on the video stalling length and frequency objective measures.

- **Number of stalls ($N[t]$):** $N[t]$ represents the number of stalls that the user experiences till time slot t.

- **Stalls length ($W[t]$):** $W[t]$ represents the length of stalls that the user experiences till time slot t.

- **Instantaneous throughput ($\mathfrak{I}[t]$):** $\mathfrak{I}[t]$ represents the instantaneous download rate obtained at every time slot t, expressed in bits per second.

- **Energy consumption ($\mathfrak{E}[t]$):** $\mathfrak{E}[t]$ represents the energy consumed to download data at every time slot t, expressed in Joules.

- **Actual streaming time ($\mathfrak{S}[t]$):** Due to the channel condition variations, the estimated rate may be different from the actual transmission rate at time slot t. The data may be downloaded in less time if the actual transmission rate is higher than the estimated. $\mathfrak{S}[t]$ represents the actual amount of time needed to download the video data at every time slot t, expressed in seconds.

4.3 QoE-Aware Traffic Splitting Optimization

This paper presents a QoE-aware resource management approach for video on demand streaming applications. Our main aim is to solve traffic splitting problem capturing the balance between user QoE, delay bounds and energy consumption for video streaming applications. To achieve high quality of experience with our target application of video streaming, we want to minimize, if not eliminate, video stalls for the users. As a result, the goal for QoE is to keep the network queue backlog from building up and causing video stalls and delays. Therefore, we aim to find the best traffic splitting solution at every time slot t minimizing the delay and stabilizing network queues while reducing the average power consumption and achieving high quality of experience.

The queue length will grow infinitely when the download rate is less than the video arrival rate; the user will then experience stalling events. The queue backlog length is thus directly related to the system parameters and channel quality such as download rates over each interface. Under queue stability, all requested bits are delivered within an acceptable limited delay experienced by the user, such that, all video chunks will be delivered within their playback deadline [100]. To ensure queue stability, decisions need to be made at every time slot t based on the current queue state and system parameters, to control the change in a function at every step. This process will allow controlling the ending value of the queue backlog size from growing infinitely.

The traffic splitting problem can be formulated as a multi-objective optimization function leading to high QoE with a controlled tradeoff in energy consumption. We use Lyapunov optimization framework from queuing theory, which also provides low computational complexity and enables real-time decisions on network transmission. The Lyapunov optimization guarantees queue stability and achieves near-optimal performance for the chosen optimization objective [12] [101].

Lyapunov-based utility functions are derived to provide solution for the multi-objective optimization providing a balance between QoE, energy consumption and delay. These utility functions are computed for the set of possible download strategies which are in our case WiFi link alone, cellular link alone, and both links simultaneously. The strategy providing the maximum utility function will be selected for transmission at time slot t.

In this section, we present: (A) the Lyapunov drift-plus-penalty optimization formulation of the multi-objective function minimizing the Lyapunov drift and cost penalty function, (B) penalty cost function capturing the balance between the power consumption and QoE in addition to queue stability, and (C) the proposed solution and utility functions derivation from the Lyapunov-based multi-objective function.

4.3.1 Lyapunov Drift-Plus-Penalty Optimization Formulation

The Lyapunov optimization considers controlling and minimizing the change in the user download queue backlog size $Q[t]$ at every time slot t resulting in a scheduling algorithm that reduces delay bounds, stabilizes the queue over time and enhances QoE.

Definition 1: The Lyapunov function is a scalar measure of the network congestion.

$$\zeta(Q[t]) = \frac{1}{2}(Q[t])^2 \tag{4.2}$$

Definition 2: The Lyapunov drift function $\Delta(Q[t])$ measures the difference in the Lyapunov function between two consecutive time slots.

$$\Delta(Q[t]) = \mathbb{E}\{\zeta(Q[t+1]) - \zeta(Q[t]) \; |Q[t]\} \tag{4.3}$$

The function grows large when the system moves towards undesirable states. Therefore, system stability is achieved by taking control actions that minimize the Lyapunov function drift function $\Delta(Q[t])$. If control decisions are made every slot t to greedily minimize $\Delta(Q[t])$, then backlogs are consistently pushed towards a lower congestion state, which maintains network stability [100][101].

Lyapunov drift-plus-penalty method is used as an extension to the base Lyapunov optimization by adding a penalty $C[t]$ term weighted by a positive coefficient $V[t]$ that determines the significance of the penalty cost function. In our case, the penalty cost function is expressed in function of power consumption and quality of experience. The Lyapunov drift-plus-penalty approach uses drift steering technique for achieving real-time near-optimal performance-delay tradeoffs for dynamic resource management [12] [101].

The Lyapunov drift-plus-penalty method is used to capture queue backlog stability in real-time network systems while optimizing the penalty objective metric which allows a balance between delay, QoE and energy in our case. The advantages of using this approach is that (1) it converges to $[O(1/V), O(V)]$ performance-delay tradeoff; it results in a time average penalty that is within $O(1/V)$ of optimality, with a corresponding $O(V)$ tradeoff in average queue size, (2) it provides local optimum guarantees even for non-convex functions, and (3) it provides simple and fast solutions by making decisions based on the current queue states and system parameters without requiring knowledge of the probabilities associated with future random events such as arrival rates and channel variation [101].

The objective function of the Lyapunov drift-plus-penalty approach will be:

$$\underset{\ell \in L[t]}{\operatorname{argmin}} \quad \Delta(Q[t]) + V[t] \cdot \mathbb{E}\{C_\ell[t] \ |Q[t]\} \tag{4.4}$$

The objective function aims at (1) minimizing the Lyapunov drift to ensure queue stability and prevent the queue to grow large, and (2) minimizing the cost function at every time slot t that is in our case expressed in terms of power consumption and quality of experience. The goal is to find the best traffic splitting strategy ℓ that minimizes the Lyapunov drift to ensure minimum delay and queue stability while reducing the transmission cost at every time slot t. Since the system is dynamic and the channel conditions and rates estimation vary over time, the positive weight $V[t]$ and cost $C_\ell[t]$ are time dependent and may vary based on the link selected for transmission at each time slot t.

For real-time traffic splitting decisions, the multi-objective function will be used to derive utility functions computed at every time slot t to choose the most efficient traffic splitting strategy ℓ among $L[t]$ represented by the following: (1) WiFi only (W), (2) cellular only (C), (3) both links simultaneously (WC), and (4) no transmission (0). These utility functions can be obtained by developing the objective function as follows.

Combining (4.1), (4.2) and (4.3), the Lyapunov drift function will be:

$$\Delta(Q[t]) = \frac{1}{2}\mathbb{E}\{(Q[t] - \mu[t] + A[t])^2 - Q[t]^2 | Q[t]\} \tag{4.5}$$

$$\leq \frac{1}{2}\mathbb{E}\{\mu[t]^2 + A[t]^2 \ |Q[t]\} - Q[t] \cdot \mathbb{E}\{\mu[t] - A[t] \ |Q[t]\} \tag{4.6}$$

Therefore, the multi-objective function in (4.4) can be upper bounded as follows:

$$\Delta(Q[t]) + V[t] \cdot \mathbb{E}\{C_\ell[t] \ |Q[t]\} \leq \frac{1}{2}\mathbb{E}\{\mu_\ell[t]^2 + A[t]^2 \ |Q[t]\} - Q[t] \cdot \mathbb{E}\{\mu_\ell[t] \ |Q[t]\} + Q[t] \cdot \mathbb{E}\{A[t] \ |Q[t]\} + V[t] \cdot \mathbb{E}\{C_\ell[t] \ |Q[t]\} \tag{4.7}$$

The amount of data $\mu_\ell[t]$ to be sent over link ℓ can be estimated at the user side based on the transmission data rate $R_\ell[t]$ representing $R_{\mathrm{W}}[t]$, $R_{\mathrm{C}}[t]$ or $R_{\mathrm{WC}}[t]$. Therefore, $\mu_\ell[t]$ is replaced by its estimate amount of data transfer when using resource management strategy ℓ at time slot t expressed as follows: $\mathbb{E}\{\mu_\ell[t] \ |R_\ell[t]\}$.

Minimizing our target multi-objective function (4.4) can thus be achieved by minimizing the upper bound of the objective function in (4.7). Define $B[t]$ and λ as follows:

$$B[t] = \frac{1}{2}\mathbb{E}\{\mu_\ell[t]^2 + A[t]^2 \ |Q[t]\} \tag{4.8}$$

$$\lambda = \mathbb{E}\{A[t] \ |Q[t]\} = \mathbb{E}\{A[t]\} \tag{4.9}$$

$B[t]$ and λ are non controllable parameters. λ represents the expected data arrival rate $A[t]$ defined by the application which is in our case the video arrival rate. λ cannot be controlled by the user and is independent of the current queue back $Q[t]$. $B[t]$ is the sum of the variances of the transmission rate and the arrival rate, which are non controllable parameters. In addition, $B[t]$ is assumed to be bounded by a fixed value B [12]. Thus, minimizing the Lyapunov drift and penalty will result in minimizing the controllable part of the upper bound in (4.7) $-Q[t] \cdot \mathbb{E}\{\mu_\ell[t] \ |Q[t]\} + V[t] \cdot \mathbb{E}\{C_\ell[t] \ |Q[t]\}$ which is equivalent to $-\mathbb{E}\{Q[t] \cdot \mu_\ell[t] - V[t] \cdot C_\ell[t] \ |Q[t]\}$.

Using the concept of opportunistically maximizing an expectation, the upper bound expression is maximized by choosing the traffic splitting strategy ℓ every time slot t as follows [101]:

$$\underset{\ell \in L[t]}{\operatorname{argmax}} \quad Q[t] \cdot \mathbb{E}\{\mu_\ell[t] \ |R_\ell[t]\} - V[t] \cdot C_\ell[t] \tag{4.10}$$

$L[t]$ is the set of possible traffic splitting decision at time slot t. The solution of the optimization problem is to find the best download strategy ℓ providing the highest performance gains as the best balance between QoE, delay, and energy consumption. The selected strategy ℓ will determine the amount of data $\mu_\ell[t]$ to be downloaded during time slot t as $\mu_{\mathrm{W}}[t]$ if WiFi link only is selected, $\mu_{\mathrm{C}}[t]$ if cellular link only, and $\mu_{\mathrm{WC}}[t]$ if both links are used simultaneously.

4.3.2 Cost Function in Terms of Power Consumption and QoE Metric For Video Streaming

We define the cost $C_\ell[t]$ of using a transmission link ℓ at time slot t as a function of power consumption and QoE parameters as follows:

$$C_\ell[t] = f(P_\ell[t], \phi_\ell[t]) = w_1 P_\ell[t] - w_2 \phi_\ell[t] \tag{4.11}$$

where w_1 and w_2 are positive weights that define the relative importance of the power consumption and QoE metrics. $P_\ell[t]$ is the power expenditure at time slot t when using strategy ℓ and $\phi_\ell[t]$ is a metric representing the quality of experience.

$P_\ell[t]$ represents the average power consumed by the mobile device while receiving data over different interfaces over time slot following time instance t. When WiFi link is selected, the mobile device will consume $P_{\mathrm{W}}[t]$ to download data during time slot t. Similarly, the device will consume $P_{\mathrm{C}}[t]$ while receiving data over cellular link only. When both links are used simultaneously for data download, the device will use both interfaces in parallel and will consume $P_{\mathrm{WC}}[t] = P_{\mathrm{W}}[t] + P_{\mathrm{C}}[t]$.

The quality of experience metric $\phi_\ell[t]$ is based on both objective and subjective psychological measures of using an information and communication technology service [102][103][104]. Several factors affect quality of the video experienced by the user end such as: (1) network parameters including transmission rate, packet loss, delay, and jitter resulting in stalls and freeze frames (2) application type and characteristics, for instance, video characteristics such as size, frame rate and resolution, and (3) user characteristics such as user's age, and interests. The satisfaction of the user when using the application can be measured by Mean Opinion Score (MOS) [105][106][107]. The MOS ranges between 1 (bad) and 5 (excellent) [102].

For video on demand streaming, the video bit rate, frame rate, compression parameters, codec and resolution are non-adaptive and fixed. Video streaming is characterized by playing synchronized media streams in a continuous way while those streams are being downloaded from the application server without having to wait for the entire video to be delivered. Once the playout phase starts, the player fetches video frame from the buffer at a constant speed defined by the video characteristics. When the service transmission rate is less than the arrival data rate, the playing buffer becomes empty. In this case the player pauses and the user will experience stalling and re-buffering events. The video streaming data is not lost, instead the frames are delayed, rather than being skipped. The last received frame freezes and is displayed until the data for the next frame is being downloaded. Therefore, the main distraction for the user and quality satisfaction degradation factors are stalling events, frequency and length. In our model, compression artifacts and losses impairments are not considered since the compression rate is non adaptive and the frame is only displayed when its data is completely downloaded. As a result, common QoE metrics relying on frame by frame and pixel analysis such as peak-signal-to-noise ratio (PSNR) [108], structural similarity (SSIM) [109] and video quality metric (VQM) [104] are not suitable for our model and cannot be considered [110].

In our work, we estimated the MOS values based on a QoE metric that is derived from standards to better capture the video quality in our optimization and performance assessment. We used re-buffering artifact QoE metric presented in Recommendation ITU-T P.1201 (2013) considering stalling and initial buffering for several reasons: (1) it assesses the effect of perceptual buffering-related indicator to the overall media session quality score, (2) the model predicts mean opinion scores on a 5-point scale as defined in ITU-T P.911, (3) it does not consider the effects of audio level, noise, delay and other impairments related to the payload, and (4) it can be applied for non-adaptive, progressive download type media streaming such as YouTube and operator managed video services over Transmission Control Protocol (TCP) [111]. For these reasons, ITU-T P.1201 QoE metric is found to be valid and suitable to our model and consistent with our considerations. In addition, the ITU-T P.1201 QoE metric $\Phi_\ell[t]$ can be customized to make real time decisions every time slot t. $\Phi_\ell[t]$, presented in [111], can be expressed as follows:

$$\Phi_\ell[t] = 5 - \max(\min((\Omega_\ell[t] + \Gamma_\ell), 4), 0) \tag{4.12}$$

where $\Omega_\ell[t]$ and Γ_ℓ are the expected degradation caused by stalls and initial buffering till time t, respectively, when using link ℓ. They are defined as follows:

$$\Omega_\ell[t] = \max(\min(s_4 + s_1 \cdot \exp((s_2 \cdot L_\ell[t] + s_3)N_\ell[t]), 4), 0) \tag{4.13}$$

$$\Gamma_\ell = \begin{cases} \max(\min(d_1 \cdot \log(T_0 + d_2), 4), 0), & \text{if } T_0 \geq 1 - d_2 \\ 0, & \text{otherwise} \end{cases} \tag{4.14}$$

where T_0 is the initial loading time in seconds, $L_\ell[t]$ is the averaged stalling duration in seconds and $N_\ell[t]$ is the number of stalling events excluding initial buffering happening till time slot t when using link ℓ. The coefficients s_1, s_2, s_3, s_4, d_1 and d_2 have the following values -1.72, -0.04, -0.36, 1.66, 0.29 and -3.29, respectively [111].

In our model, we assume the QoE is not affected by the initial buffering degradation; the video is either played without initial buffering or the initial buffering is lower than 4.71 $(1-d_2)$ seconds. Accordingly, the main degradation of the QoE is caused by the re-buffering and stalling artifacts. To study the effect of

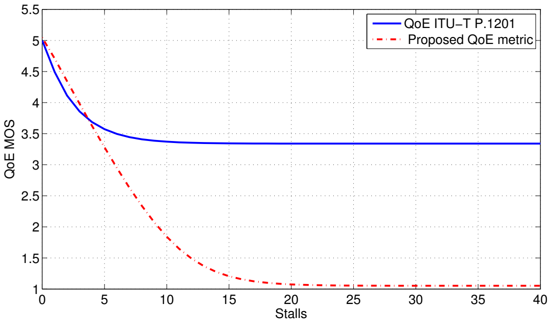


Figure 4.2: QoE ITU-T P.1201 metric and proposed QoE metric variations with respect to the number of stalls considering average stalling length of 1 second.

the lengths and frequency of occurrence of stalling events on the QoE, Figure 4.2 shows the QoE MOS while varying the number of stalls. In the considered case scenario, the average stalling length is considered 1 second. The results show that the QoE varies from 5 (excellent) where the user experience no stalls, to 3.34 where the number of stalls is high. The QoE values are limited to 3.34 since the initial buffering artifacts are not considered. In addition, the re-buffering degradation score $\Omega_\ell[t]$, expressed in (4.13), ranges between 0 and 1.66.

In general, the MOS needs to vary between 1 and 5, however, the QoE metric presented by ITU-T P.1201 is limited to 3.34. For this reason, we utilize an updated metric inspired from the ITU-T P.1201 QoE metric to better emphasize the impact of stalls on MOS. To this end, we use a logarithmic transformation curve fitting to transform the ITU-T P.1201 QoE metric scale from 3.34-5 to 1-5. Accordingly, the new metric $\phi_\ell[t]$ can be obtained from the ITU-T P.1201 QoE metric $\Phi_\ell[t]$ as follows:

$$\phi_\ell[t] = a \cdot \log(b \cdot \Phi_\ell[t] + c) \tag{4.15}$$

where a, b and c are found to be 0.9377, 128.9 and -427.6, respectively. As shown in Figure 4.2, the proposed QoE metric values range between 1 and 5. For instance, when the number of stalls is 9 and 16, the MOS score provided using ITU-T P.1201 QoE metric was 3.3715 and 3.3429, respectively. The video is then considered fair, perceptible but not annoying for both cases. However, when using the proposed QoE metric in (4.15), the scores were 2.0741 and 1.1590, respectively. The results emphasized the impact of stalls; accordingly, when the user experience 16 stalls, the video will be bad and very annoying, and in case of 9 stalls, the video will be poor and annoying.

The QoE metric $\phi_\ell[t]$ will be integrated in the Lyapunov drift-plus-penalty objective to capture QoE-related tradeoffs.

4.3.3 Solution Approach: Traffic Splitting with delay-power-QoE Balance

The Lyapunov drift-plus-penalty problem formulated in (4.10) can be developed by expressing the cost function in terms of power consumption and QoE. The weighted cost function in (4.11) will be:

$$\begin{aligned} V[t] \cdot C_\ell[t] &= V[t]\,(w_1 \cdot P_\ell[t] - w_2 \cdot \phi_\ell[t]) && (4.16)\\ &= V_1[t] \cdot P_\ell[t] - V_2[t] \cdot \phi_\ell[t] && (4.17)\end{aligned}$$

Therefore, the objective function will be:

$$\underset{\ell \in L[t]}{\operatorname{argmax}} \qquad Q[t] \cdot \mathbb{E}\{\mu_\ell[t]\ |R_\ell[t]\} - V_1[t] \cdot P_\ell[t] + V_2[t] \cdot \phi_\ell[t] \tag{4.18}$$

where $V_1[t]$ and $V_2[t]$ are positive weights that define the relative importance of the power and QoE metrics in the objective function.

In our problem, the decision variable is the strategy ℓ to be selected at time slot t providing the best performance. One of four possible download strategies can be selected. The decision will be either selecting WiFi link only, cellular link only, both links simultaneously or no transmission at time slot t. Therefore, the solution of the problem can be achieved by computing the following utility function for the possible resource management strategies considered at every time slot t. The optimal solution is given by the decision that maximizes the utility per time slot [101].

$$U_\ell[t] = Q[t] \cdot \mathbb{E}\{\mu_\ell[t]\ |R_\ell[t]\} - V_1[t] \cdot P_\ell[t] + V_2[t] \cdot \phi_\ell[t] \tag{4.19}$$

$U_\ell[t]$ is computed for different download strategies as follows:

$$\begin{aligned} U_{\mathrm{W}}[t] &= Q[t] \cdot \mathbb{E}\{\mu_{\mathrm{W}}[t]\ |R_{\mathrm{W}}[t]\} - V_1[t] \cdot P_{\mathrm{W}}[t] + V_2[t] \cdot \phi_{\mathrm{W}}[t] && (4.20)\\ U_{\mathrm{C}}[t] &= Q[t] \cdot \mathbb{E}\{\mu_{\mathrm{C}}[t]\ |R_{\mathrm{C}}[t]\} - V_1[t] \cdot P_{\mathrm{C}}[t] + V_2[t] \cdot \phi_{\mathrm{C}}[t] && (4.21)\\ U_{\mathrm{WC}}[t] &= Q[t] \cdot (\mathbb{E}\{\mu_{\mathrm{W}}[t]\ |R_{\mathrm{W}}[t]\} + \mathbb{E}\{\mu_{\mathrm{C}}[t]\ |R_{\mathrm{C}}[t]\})\\ &\quad - V_1[t] \cdot (P_{\mathrm{W}}[t] + P_{\mathrm{C}}[t]) + V_2[t] \cdot \phi_{\mathrm{WC}}[t] && (4.22)\end{aligned}$$

where $U_{\mathrm{W}}[t]$, $U_{\mathrm{C}}[t]$ and $U_{\mathrm{WC}}[t]$ are the utility functions of using WiFi link alone, cellular link alone, and both links simultaneously, respectively, at time slot t. The strategy providing the maximum utility function will be selected for transmission at time slot t [101]. However, when the three utility functions are negative, there is no benefit of sending over the links since the device will be consuming more power than benefiting from downloading the data in terms of throughput and QoE; no transmission is recommended in this case. The proposed TS-PQ approach performs in real-time, autonomously at the user end following the steps shown in Figure 4.3. The video specifications such as video size, duration and frame rate,

Algorithm 4.1: The proposed multi-objective traffic splitting with delay-power-QoE balance TS-PQ

```
Input:
- Video specifications: video size, duration, frame rate, arrival rate $A[t]$
- Time slot duration: $T_\mathrm{s}$
- Available interfaces (cellular/WiFi) and set of possible download strategies solutions: $L[t]$
- Power consumption and QoE metric weights: $V_1[t]$ and $V_2[t]$, respectively
- Initial queue backlog size: $Q[0] = 0$
- Initial video data downloaded: $\mu[0] = 0$
- Initial video data played: $Y[0] = 0$
- Initial video data downloaded but not yet played: $D[0] = 0$
- Initial number of stalls: $N[0] = 0$
- Initial stalls length: $W[0] = 0$

Output:
The download strategy $\ell[t] \in L[t] = \{0, \mathrm{W}, \mathrm{C}, \mathrm{WC}\}$

Estimate WiFi and cellular transmission rates $R_\mathrm{W}[t]$ and $R_\mathrm{C}[t]$, respectively, by dividing the data downloaded over the time required for download
Estimate next chunk data size using each strategy $\ell$, $\mathbb{E}\{\mu_\ell[t]\ |R_\ell[t]\}$: (1) $\mathbb{E}\{\mu_\mathrm{W}[t]\ |R_\mathrm{W}[t]\}$, (2) $\mathbb{E}\{\mu_\mathrm{C}[t]\ |R_\mathrm{C}[t]\}$, and (3) $\mathbb{E}\{\mu_\mathrm{WC}[t]\ |R_\mathrm{WC}[t]\}$ as follows: $\mathbb{E}\{\mu_\ell[t]\ |R_\ell[t]\} = R_\ell[t] \cdot T_\mathrm{s}$
Estimate quality of experience $\phi_\ell[t]$ based on (4.15) for every strategy: $\phi_\mathrm{W}[t]$, $\phi_\mathrm{C}[t]$, $\phi_\mathrm{WC}[t]$ by estimating the number of stalls $N_\ell[t]$ and average length of stalls $L_\ell[t]$ as follows:
    • if $D[t-1] + \mathbb{E}\{\mu_\ell[t]\ |R_\ell[t]\} < A[t]$ then
    •   Update $N_\ell[t] = N[t-1] + 1$
    •   Estimate stalling length in time slot $t$ as $\frac{A[t]-(D[t-1]+\mathbb{E}\{\mu_\ell[t]\ |R_\ell[t]\})}{A[t]} \cdot T_\mathrm{s}$
    •   Compute $L_\ell[t] = \frac{W[t-1]+\frac{A[t]-(D[t-1]+\mathbb{E}\{\mu_\ell[t]\ |R_\ell[t]\})}{A[t]} \cdot T_\mathrm{s}}{N_\ell[t]}$
    • else
    •   Update $N_\ell[t] = N[t-1]$
    •   Compute $L_\ell[t] = \frac{W[t-1]}{N_\ell[t]}$
    • end if
Compute the utility functions $U_\mathrm{W}$, $U_\mathrm{C}$ and $U_\mathrm{WC}$ as described in (4.20), (4.21), and (4.22)
Select the strategy $\ell$ providing the higher utility function. No transmission when utility functions are all negative
Send request to download the chunk from the server
Compute the actual data downloaded $\mu[t]$ and time needed to download the data
Update queue backlog $Q[t] = Q[t-1] - \mu[t] + A[t]$
Update data played $Y[t] = \min(D[t-1] + \mu[t], A[t])$
Update data downloaded but not played $D[t] = \max(\mu[t] - Y[t], 0)$
Update QoE $\phi[t]$ based on (4.15), number of stalls $N[t]$, stalls length $W[t]$, instantaneous throughput $\mathfrak{I}[t]$, and energy consumption $\mathfrak{E}[t]$
```

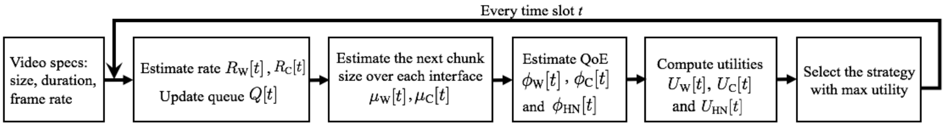

Every time slot t
Video specs: size, duration, frame rate
Estimate rate $R_{\mathrm{W}}[t]$, $R_{\mathrm{C}}[t]$
Update queue $Q[t]$
Estimate the next chunk size over each interface
$\mu_{\mathrm{W}}[t]$, $\mu_{\mathrm{C}}[t]$
Estimate QoE
$\phi_{\mathrm{W}}[t]$, $\phi_{\mathrm{C}}[t]$
and $\phi_{\mathrm{HN}}[t]$
Compute utilities
$U_{\mathrm{W}}[t]$, $U_{\mathrm{C}}[t]$
and $U_{\mathrm{HN}}[t]$
Select the strategy with max utility

Figure 4.3: Diagram representing the proposed TS-PQ approach for near real-time optimized traffic selection.

are obtained from the server before the start of the video download. The cycle starts with estimating the rate provided by each link and updating the queue based on the arrival rate and data received. Based on the rate estimation, the data for the next time slot is estimated, and is considered for data download over each interface. The quality of experience is estimated over each link using (4.15) based on the estimated number of stalls and length over each interface. The utility for each strategy is then computed using (4.20), (4.21), and (4.22). The strategy providing the maximum overall utility function is selected at every time slot t for data download. The cycle is then repeated after updating statistics such as transmission time, data rates, queue length, number and duration of stalls, until video data is downloaded. The steps of the method are shown in Algorithm 4.1 along with the needed parameters and computations for the proposed approach.

In regards to complexity of our method, the proposed approach is scalable since the traffic splitting decision is made independently at the user end. The method can accommodate for multi-users and multi-interfaces. In the case of multi-user scenario, every user is responsible for her or his own decisions based on its system parameters. If n interfaces are available for a user, $(2^n - 1)$ utility functions are computed. In our case study, we consider the coexistence of WiFi and cellular networks. For this case, as previously described, three utility functions are computed at each user end: $U_{\mathrm{W}}[t]$, $U_{\mathrm{C}}[t]$ and $U_{\mathrm{WC}}[t]$ corresponding to the following transmission strategies: WiFi only, cellular only and both interfaces simultaneously, respectively.

4.4 Results and Discussion

To validate the proposed QoE-aware traffic splitting approach under realistic conditions, experimental measurements are used to determine WiFi and cellular key link parameters, such as effective download rate and energy consumed per second during data reception. The obtained link parameter values are then used to quantify and analyze the performance of the proposed QoE-aware Lyapunov-based approach for HetNet resource management.

4.4.1 Experimental Energy Measurements

Experimental measurements were conducted to capture the effect of signal strength and traffic load on the effective download rate and energy consumption. An Android application was developed on the Samsung Galaxy SIII device to download data of different sizes ranging between 100 KB and 1.2 MB from an HTTP server via WiFi (802.11b) and 3G/4G cellular links. The collection was repeated at different locations where we could have varying traffic loads and signal strengths, which included home (low load) and library (high load) network environments, and different signal strengths such as close and far from the WiFi access point, indoor and outdoor scenarios. In each scenario, the data rate was obtained from the application while power consumption was measured using a data acquisition device from National Instruments monitored via a LABVIEW application. The results showed that the mobile device consumes more energy when receiving on the 3G interface than when receiving on the WiFi interface. The average power consumed was 1.307 Watts for WiFi and 1.859 Watts for 3G.

4.4.2 Performance Evaluation

In order to assess the performance effectiveness of the proposed cellular/WiFi traffic splitting approach, we generated results for the following different strategies including state-of-the-art related work from the literature:

1. *WiFi only* (WO): the user downloads data using WiFi link only.

2. *Cellular only* (CO): the user downloads data using WiFi link only.

3. *Maximum rate network selection* (MaxR-NS): the link providing the higher rate is selected in every time slot.

4. *Minimum energy network selection* (MinE-NS): the link proving the lower energy consumption is selected in every time slot.

5. *Stable and adaptive link selection approach* (SALSA): the network providing higher delay-power tradeoff utility function is selected. The authors in [12] presented a network selection approach for energy-delay tradeoff using the Lyapunov optimization framework for video upload. The weight $V[t]$ of the power metric is considered to be variable over time to adapt the impact of power based on the queue size and delay. We compare our proposed approach to SALSA since it also uses Lyapunov drift-plus-penalty for HetNet resource management, however, our approach is different since it considers traffic splitting in addition to quality of experience.

6. *Traffic splitting using both links simultaneously* (TS-S): the user always uses both links simultaneously to download data.

7. *Traffic splitting with delay-power balance* (TS-P): the user uses the strategy that provides higher utility based on our proposed delay-power tradeoff, with fixed weighing factor $V[t]$ of the power metric and considering traffic splitting option. Therefore, the cost function in this case only captures power without considering QoE. The objective function (4.18) is reduced to:

$$\underset{\ell \in L[t]}{\operatorname{argmax}} \quad Q[t] \cdot \mathbb{E}\{\mu_\ell[t] \; |S_\ell[t], R_\ell[t]\} - V[t] \cdot P_\ell[t] \tag{4.23}$$

   where the queue length $Q[t]$ and $V[t]$ give weights to the transmission rate and power consumption respectively.

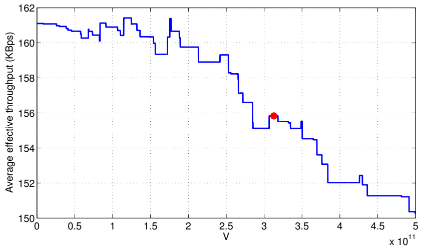
8. *Traffic splitting with delay-power-QoE balance TS-PQ*: the user uses the strategy including traffic splitting that provides higher utility based on our proposed QoE-aware resource management approach providing delay-cost tradeoff where cost is a function of both power and QoE with weighing factors $V_1[t]$ and $V_2[t]$, respectively.

4.4.3 Simulations Setup

We first evaluated the performance of our proposed approach using simulations conducted using MATLAB to stream a video using different strategies. We used MATLAB for convenience, which allowed us to easily access and modify the physical layer and provided high flexibility in modifying the data requests.

In our model, a mobile device downloads data for video streaming application, where the video specifications, such as video size, duration, and frame rate, are obtained as input from the server before the start of the video download. The chosen video has a size of 7 MBytes, a duration of 60 seconds, a frame rate of 25 fps, and an arrival rate of 117 KBytes every second. The cellular and WiFi transmission rates were assumed to have exponential distribution with different mean values as presented in the results section below.

As presented in Algorithm 4.1 and Figure 4.3, at each time slot of duration 1 second, the transmission rate, queue size, QoE and power consumption are first estimated. The quality of experience is estimated over each link using equation (4.15) based on the estimated number of stalls and length over each interface. The mobile device makes decision on the link combination that provides the highest utility function, where the options are: (1) WiFi only, (2) cellular only, (3) both links simultaneously, or (4) no transmission. Once data is downloaded based on the selected strategy, the actual parameters are recorded such as queue size, transmission rate, QoE and energy consumption. Recordings become part of the inputs for estimating the parameters of the next time slot. The queue is updated based on the arrival rate and data received. If the transmission rate is lower than the video arrival rate, the mobile device is able to download only a fraction of the requested data. The remaining data that was not downloaded is stored in



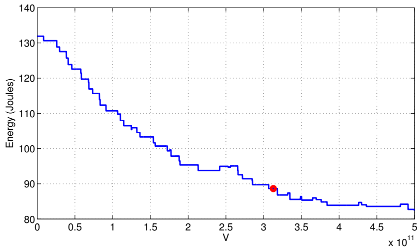


Figure 4.4: Average effective throughput in KBps variation with respect to V. The highlighted value V, equal to $3.128 \cdot 10^{11}$, represents a tradeoff between energy and throughput.

Figure 4.5: Energy consumption in Joules variation with respect to V. The highlighted value V, equal to $3.128 \cdot 10^{11}$, represents a tradeoff between energy and throughput.

the queue to be downloaded in the next time slots. If the transmission rate is higher than the video arrival rate, the data is downloaded on time without any delay, stalls, or freeze frames. In this case, the queue is empty with a queue size of zero. The process is then repeated, statistics are updated every time slot until video data is completely downloaded.

4.4.4 Simulations Results and Analysis

To compare the performance of the various strategies mentioned in Section 4.4.2, we evaluated the queue size, the average throughput, total energy consumption, delay and QoE for three case scenarios. In the first two case scenarios, symmetric rate for both WiFi and cellular is considered. In the first case, WiFi and cellular transmission rate follow exponential distribution with average rate of 110 KBps. In the second case, WiFi and cellular transmission rate follow exponential distribution with average rate of 450 KBps. In the third scenario, we considered asymmetric rates; we fixed the value of the average WiFi rate of 200 KBps and varied the average cellular rate from 1 KBps to 600 KBps. In this section, we present analysis for the selected power and QoE weights used in our simulations and performance results for the considered scenarios. In addition, we present a study on the duration of time slot and its effect on the performance of the considered approaches.

Study on the Power Consumption and QoE Weights

As presented in (4.23), when TS-P is used, the queue length $Q[t]$ and $V[t]$ give weights to the transmission rate and power consumption, respectively. Accordingly, when the queue length of unfinished work is high, the transmission rate will have more impact than power and the transmission will occur even if the trans-

Table 4.1: Simulations results and statistics with R_{W} and R_{C} average of 110 KBps

	WO	**CO**	**MaxR-NS**	**MinE-NS**	**SALSA**	**TS-S**	**TS-P**	**TS-PQ**
Total streaming time (s)	72.4	68.64	64.64	64.76	64.64	62.04	63.2	63.2
Total delay (s)	12.4	8.64	4.64	4.76	4.64	2.04	3.2	3.2
Average throughput (KBps)	98.9	104.3	110.8	110.6	110.8	115.4	113.3	113.3
Average eff. throughput (KBps)	114.7	108.4	147.1	142.6	147.1	216.2	155.8	155.9
Total energy consumption (J)	88.2	126.3	85.8	83.9	85.8	143.1	88.6	90.4
Average queue size (KB)	532.1	394.6	116.6	132.3	116.6	37.3	80.3	76.82
Maximum queue size (KB)	1430.2	972.2	536.4	547.8	536.4	231.8	367.1	367.1
Number of stalls	23	21	12	12	12	7	8	7
Number of freeze frames	310	216	116	119	116	51	80	80
QoE ϕ (4.15)	1.01	1.03	1.57	1.57	1.57	2.77	2.46	2.74

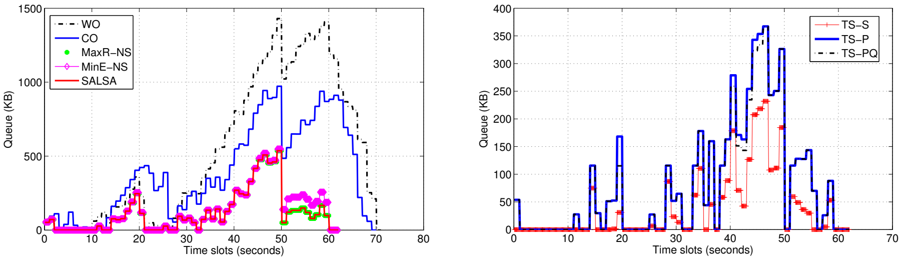
mission rate is too small. However, when the queue length is low, the power has more impact; thus, the approach can decide to deffer transmission to save energy. To give the power and rate similar impact, the value $V[t]$ should be estimated based on the observed values for queue length, transmission rate, and power consumption through simulations. Figures 4.4 and 4.5 show the average effective throughput and energy consumption variations with respect to $V[t]$. In this considered scenario, the WiFi and cellular transmission rate followed an exponential distribution with average of 110 KBps. To obtain a tradeoff between throughout and energy, we selected the value of V highlighted in red to be $3.128 \cdot 10^{11}$. Similarly, the value of $V[t]$ was selected to be $6.256 \cdot 10^{11}$ when the average WiFi and cellular transmission rate was set to 450 KBps.

Using TS-PQ, $V_1[t]$ and $V_2[t]$ present the weights for power consumption and QoE, respectively. Using the same analysis, $V_1[t]$ was chosen to be $3.128 \cdot 10^{11}$ and $6.256 \cdot 10^{11}$ when the average link transmission rate was set to 110 KBps and 400 KBps, respectively. Similarly, $V_2[t]$ is fixed to be $5.52 \cdot 10^{10}$ and $11.04 \cdot 10^{12}$ when the average link transmission rate was set to 110 KBps and 450 KBps, respectively.

Results for Scenario 1: WiFi and Cellular Average Rate of 110 KBps

In the first scenario, we considered a symmetric rate for WiFi and cellular links. The transmission rates are modeled following an exponential distribution with average rate of 110 KBps. Table 4.1 presents the results for scenario 1 and compares the performance of the various approaches presented in Section 4.4.2 in terms of (1) total streaming time which is the time required to stream and play all the video, (2) delay which is the total duration of stalls, (3) average throughput which is computed by dividing the total amount of data received by the total streaming time, (4) average effective throughput which computed by averaging the throughput achieved in every time slot t, (5) total energy consumed for downloading the video, (6) average and maximum queue size, (7) number of stalls which represents to rebuffering events experienced by the user, (8) number of freeze frames, and (9) QoE MOS computed based on (4.15).

The results show low performance for WiFi only, cellular only and network



(a) Network selection approaches

(b) Traffic splitting approaches

Figure 4.6: Queue size in KB variation over time with R_{W} and R_{C} average of 110 KBps.

selection strategies. The user will experience more delay, high queue length, freeze frames and stalls when one link is selected instead of traffic splitting on both links. The proposed TS-PQ approach provided the best balance in terms of throughput, queue stability, energy consumption and user satisfaction.

To show performance in terms of queue stability, Figure 4.6 shows the queue length variation over time. The queue size varies based on the service transmission rate and the arrival data process. The queue size will go very large if the transmission rate is lower than the arrival data rate. Otherwise, the data will be downloaded on time and the queue size will be 0. The results in Figure 4.6 show that traffic splitting approaches TS-S, TS-P and TS-PQ provided higher queue stability, lower queue length and lower delay. TS-S approach showed higher queue stability since the data is always downloaded simultaneously over both links.

To quantify the tradeoff between QoE and energy consumption, Figure 4.7 presents the total energy consumption versus the QoE mean opinion score for each approach. In addition, Figure 4.8 presents the tradeoff between total energy consumption in Joules versus the average effective throughput in KBps. The aim is to minimize the energy expenditure while increasing user quality of experience and throughput. The areas with best performance are highlighted in each figure; these areas lead to high QoE and throughput with reduced energy consumption.

The approaches where traffic splitting is considered provided the best performance in terms of delay, queue length, throughput and user satisfaction while consuming very high energy. TS-P consumed lower energy consumption with a performance reduction in throughput, delay and QoE since it aims to provide a tradeoff between energy-throughput without considering QoE. Our proposed approach TS-PQ aimed to provide a balance between QoE, energy consumption and delay. The results for scenario 1 showed that TS-PQ provided QoE of 2.74 which is higher than the QoE provided by TS-P (2.46) and close to the QoE provided by the TS-S approach (2.77). However, the proposed approach TS-PQ provided

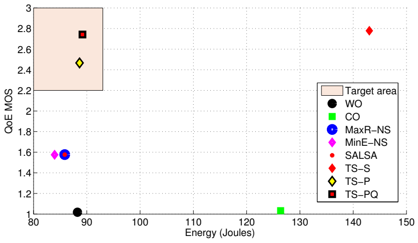


Figure 4.7: Total energy expenditure (J) and QoE MOS for every approach with R_W and R_C average of 110 KBps.

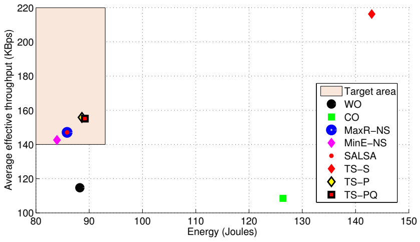


Figure 4.8: Total energy expenditure (J) and average effective throughput (KBps) for every approach with R_W and R_C average of 110 KBps.

lower energy cost of 89.1 Joules which is 37.7% less than the energy consumed when using both links simultaneously TS-S. The number of freeze frames and stalls were the lowest with 7 stalls and 80 freeze frames which leads to a delay of 3.2 seconds.

Considering the same scenario 1 where the WiFi and cellular rates are symmetric and modeled following an exponential distribution with average rate of 110 KBps, initial buffering is considered before starting video playout. We assumed the video playout starts after downloading specific amount of data corresponding to 2 seconds of video streaming. Table 4.2 presents the performance results for scenario 1 with initial buffering considerations and compares the performance of the various approaches presented in Section 4.4.2 in terms of (1) total streaming time which is the time required to stream and play all the video, (2) delay which is the total duration of stalls, (3) average throughput which is computed by dividing the total amount of data received by the total streaming time, (4) average effective throughput which computed by averaging the throughput achieved in every time slot t, (5) total energy consumed for downloading the video, (6) average and maximum queue size, (7) number of stalls which represents to rebuffering events experienced by the user, (8) number of freeze frames, and (9) QoE MOS computed based on equation (4.15). As presented in Table 4.2, adding initial buffering enhanced the quality of experience of all the compared strategies. The number of stalls is reduced since the initial buffer allowed to download more data before starting the video streaming playout, which enhanced the QoE metric. The network selection strategies were able to provide the user with higher QoE of 2.75 when initial buffering is considered instead of 1.57 without initial buffering. Similarly, the TS-PQ proposed approach was able to provide higher QoE of 3.37. The initial buffering does not have an effect on the transmission parameters and performance in terms of download time, throughput and energy consumption of the algorithms. The mobile is able to download data within the same duration, however, the playout is affected. Initial buffering allows the video

Table 4.2: Simulations results and statistics with R_{W} and R_{C} average of 110 KBps with initial buffering of 2 seconds of video data

	WO	**CO**	**MaxR-NS**	**MinE-NS**	**SALSA**	**TS-S**	**TS-P**	**TS-PQ**
Total streaming time (s)	72.4	68.64	64.64	64.76	64.64	62.04	63.2	63.2
Total delay (s)	12.4	8.64	4.64	4.76	4.64	2.04	3.2	3.2
Average throughput (KBps)	98.9	104.3	110.8	110.6	110.8	115.4	113.3	113.3
Average eff. throughput (KBps)	114.7	108.4	147.1	142.6	147.1	216.2	155.8	156.2
Total energy consumption (J)	88.2	126.3	85.8	83.9	85.8	143.1	88.6	89.1
Average queue size (KB)	532.1	394.6	116.6	132.3	116.6	37.3	80.3	75.7
Maximum queue size (KB)	1430.2	972.2	536.4	547.8	536.4	231.8	367.1	367.1
Number of stalls	17	15	7	7	7	4	6	5
Number of freeze frames	260	166	66	69	66	26	55	54
QoE ϕ (4.15)	1.1	1.22	2.75	2.75	2.75	3.72	3.06	3.37

Table 4.3: Simulations results and statistics with R_{W} and R_{C} average of 450 KBps

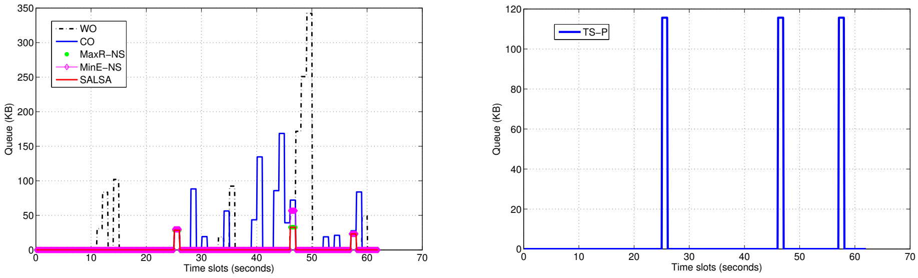
	WO	**CO**	**MaxR-NS**	**MinE-NS**	**SALSA**	**TS-S**	**TS-P**	**TS-PQ**
Total streaming time (s)	62.96	61.48	60.32	60.52	60.32	60	61	60
Total delay (s)	2.96	1.48	0.32	0.52	0.32	0	1	0
Average throughput (KBps)	111.1	113.7	115.9	115.5	115.9	116.5	114.6	116.5
Average eff. throughput (KBps)	469.1	448.1	655.1	638.3	655.1	917.3	634.5	642.5
Total energy consumption (J)	42.5	56.3	31.8	30.2	31.8	44.2	26.7	32.4
Average queue size (KB)	20.1	14.3	1.3	1.7	1.3	0	5.6	0
Maximum queue size (KB)	342.1	168.6	32.7	56.7	32.7	0	115.6	0
Number of stalls	6	4	2	2	2	0	1	0
Number of freeze frames	74	37	8	13	8	0	25	0
QoE ϕ (4.15)	3.04	3.71	4.40	4.39	4.40	5	4.70	5

playout to start after downloading 2 seconds of video, which results in a reduction in the number of freeze frames and stalls duration. This, in turn, directly affects the quality perceived by the end user and enhance the QoE.

Results for Scenario 2: WiFi and Cellular Average Rate of 450 KBps

In the second scenario, we considered WiFi and cellular links with transmission rates following an exponential distribution with average rate of 450 KBps. Table 4.3 presents the performance results for scenario 2 and compares the performance of the various approaches and parameters presented in Section 4.4.2. The results in Table 4.3 show that traffic splitting approaches provide higher performance with lower delay, freeze frames and stalls. In addition, all the approaches were able to provide better performance when the transmission rates increased from 110 KBps in scenario 1 (see Table 4.1) to 450 KBps in scenario 2 (see Table 4.3). For instance, when WiFi is only used, the delay was reduced from 12.4 in scenario 1 where the average rate is 110 KBps to 2.96 seconds where the average rate is 450 KBps; the number of stalls was reduced from 23 (scenario 1) to 6 (scenario 2). The proposed approach showed an excellent performance in terms of QoE without any stalls or buffering events similar to the performance of using both links simultaneously TS-S while consuming less energy.

Figure 4.9 shows the queue length variation over time when the WiFi and cellular average transmission rates are 450 KBps. Same analysis obtained from Fig-



(a) Network selection approaches

(b) Traffic splitting approaches

Figure 4.9: Queue size in KB variation over time with R_{W} and R_{C} average of 450 KBps.

ure 4.6 can be drawn, except that when the transmission rate is higher, more data can be downloaded which makes the queue length smaller. The proposed traffic splitting approach TS-PQ and using both links simultaneously TS-S showed an empty queue, which indicates an excellent quality of experience. The provided transmission rate was higher than the arrival rate. All the data is downloaded on time without any stalls or delay which reflects an excellent QoE performance. However, the the traffic splitting with delay-power balance TS-P approach causes stalls as shown in Figure 4.9(b).These results can also be reflected in Table 4.3.

Figures 4.10 and 4.11 show QoE-energy consumption and throughput-energy consumption tradeoffs, respectively. The proposed approach was able to perform perfectly without any delay, stalls or freeze frames and provide the user with excellent quality of experience MOS of 5. Similar performance analysis can be remarked when evaluating QoE by metrics presented in [112] and [113]. Using both links simultaneously always (TS-S) approach was able to provide similar performance in terms of quality of experience, however, the proposed approach consumes 26.6% less energy.

Adding initial buffering enhanced the quality of experience of all the compared strategies. Table 4.4 presents the results for scenario 2 with initial buffering considerations and compares the performance of the various approaches presented in Section 4.4.2. We assumed the video playout will start after downloading specific amount of data corresponding to 2 seconds of video streaming. The compared strategies showed higher user quality of experience. The network selection and traffic splitting strategies were able to provide QoE of 5 without any stalls or buffering events.

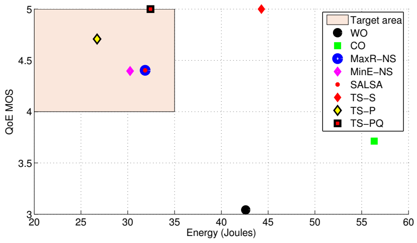


Figure 4.10: Total energy expenditure (J) and QoE MOS for every approach with R_{W} and R_{C} average of 450 KBps.

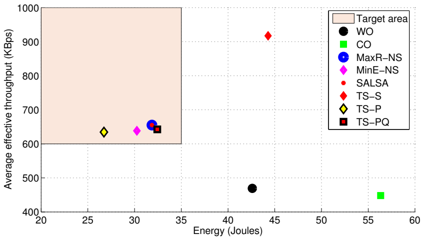


Figure 4.11: Total energy expenditure (J) and average effective throughput (KBps) for every approach with R_{W} and R_{C} average of 450 KBps.

Table 4.4: Simulations results and statistics with R_{W} and R_{C} average of 450 KBps with initial buffering of 2 seconds of video data

	WO	CO	MaxR-NS	MinE-NS	SALSA	TS-S	TS-P	TS-PQ
Total streaming time (s)	62.96	61.48	61	61	61	61	62	61
Total delay (s)	2.96	1.48	1	1	1	1	2	1
Average throughput (KBps)	111.1	113.7	114.6	114.6	114.6	114.6	112.8	114.6
Average eff. throughput (KBps)	469.1	448.1	655.1	638.3	655.1	917.3	634.5	642.5
Total energy consumption (J)	42.5	56.3	31.8	30.2	31.8	44.2	26.7	32.4
Average queue size (KB)	20.1	14.3	1.3	1.7	1.3	0	5.6	0
Maximum queue size (KB)	342.1	168.6	32.78	56.7	32.7	0	115.6	0
Number of stalls	3	2	0	0	0	0	0	0
Number of freeze frames	49	12	0	0	0	0	0	0
QoE ϕ (4.15)	4.01	4.39	5	5	5	5	5	5

Results for Scenario 3: WiFi Average Rate of 200 KBps with Different Cellular Average Rates

In the previous scenarios, we used symmetrical rate for both WiFi and cellular links. To show the performance of more realistic scenarios, we compared the performance of asymmetrical WiFi and cellular transmission rates considering video streaming for 17 hours duration which corresponds to 1020 runs of 60 seconds videos. The performance metrics such as QoE and energy consumption are measured every 60 seconds and the overall metrics represent the average over the 1020 runs. In our simulations, the WiFi transmission rate follows an exponential distribution with average of 200 KBps. The cellular rate followed the exponential distribution with average transmission rate varying from 100 KBps to 600 KBps. The performance of the proposed QoE-aware Lyapunov based approach (TS-PQ) was compared to the maximum rate network selection approach (MaxR-NS), WiFi only (WO) and cellular only (CO) approaches. Figures 4.12 and 4.13 show the performance of the mentioned approaches in terms of QoE and total energy consumption, respectively. Figure 4.12 shows the QoE MOS for the different approaches with respect to the variation of average R_{C} rate. The ITU-T P.1201 (2013) QoE metric is trained and validated for sequences having duration

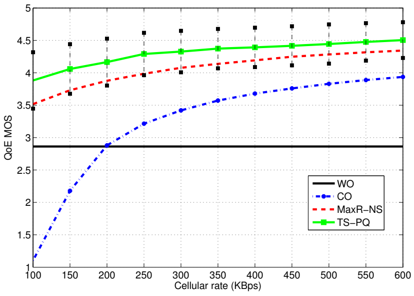


Figure 4.12: QoE MOS variation with respect to cellular average rate R_C.

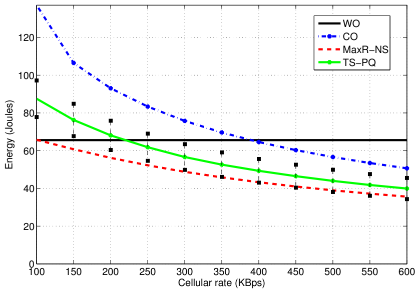


Figure 4.13: Total energy consumption in Joules with respect to cellular average rate R_C.

between 30 and 60 seconds, where the user does not interact with the player such as stop, play, rewind and fast forward. Accordingly, the QoE for 17 hours duration is measured every 60 seconds and the overall QoE metric represents the average QoE over the whole duration.

The proposed approach was able to provide a MOS of 4.5 when the average cellular rate was higher than 550 KBps, and a MOS of 3.9 when the cellular rate was 100 KBps. CO performance was enhanced with the increase of average transmission rate to achieve MOS of 2.8, similar to WO, when the WiFi and cellular rates have equal average rates, and a maximum MOS of 3.9 when the average rate is 600 KBps. The QoE performance of TS-PQ and MaxR-NS approaches also increased with the increase of the average R_C rate since the mobile device will take advantage of the better channel quality to make better decisions reducing energy while achieving high quality end-user experience.

Figure 4.13 shows that MaxR-NS provided lower energy consumption since the approach decides on the link providing higher rate every time slot. When the average R_C is low, MaxR-NS tends to select the WiFi link more often; MaxR-NS and WO have similar energy consumption. When R_W and R_C average rates are equal, WO and CO presents similar QoE, however, energy consumption is higher since mobile device consumes more power while receiving over cellular link.

TS-PQ consumed more energy than MaxR-NS and WO when the cellular average rate was small since the proposed approach will tend to use of both links simultaneously to download more data, reduce delay and maximize QoE. When the average R_C increases, TS-PQ consumes less energy since it took advantage of the intelligence of the proposed approach; the data in the queue is lower when the transmission rate is higher, thus, the impact of delay is reduced and the power has more impact on the decision.

In addition to the average values for QoE and energy consumption, the variations in each of the measures over 1020 runs (17 hours) are captured to show

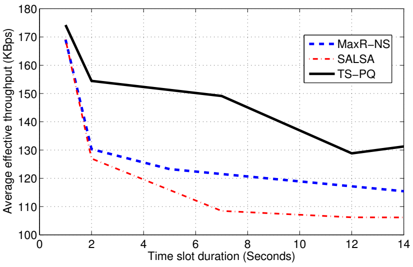


Figure 4.14: Average effective throughput in KBps variation with respect to time slot duration T_s.

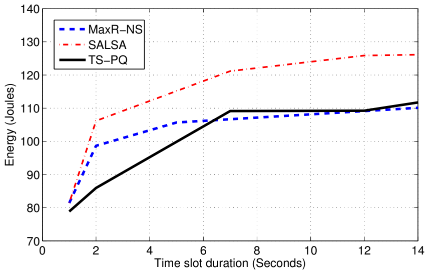


Figure 4.15: Total energy consumption in Joules variation with respect to time slot duration T_s.

the accuracy of the proposed approach. As presented in Figure 4.12, the standard deviation for QoE metric ranged between 0.2 and 0.5 which corresponds to a range of 4.44% to 14%. The standard deviation for energy consumption ranged between 5.6 and 11.2 Joules (Figure 4.13), which corresponds to a range of 10.85% to 14%. This indicates a relatively acceptable variation level in the performance of each run and validates the reliability and accuracy of our TS-PQ proposed approach.

Study on the Time Slot Duration

In our work, we used the notion of time slot to allow the feasibility of making decisions periodically. To evaluate the effect of the duration of the time slot on the performance of the compared approaches, simulations were conducted for different time slot durations varying from 1 second to 16 seconds. WiFi and cellular transmission rates follow the exponential distribution with average rate of 110 KBps. The results in Figures 4.14, 4.15 and 4.16 show the average effective throughput, energy consumption and QoE, respectively, for maximum rate network selection (MaxR-NS), SALSA and the proposed TS-PQ approach. The duration of the time slot will decide how often the transmission decision is made. In general, very small time slot duration (less than 1 second) will not be practical due to the overhead of establishing connection between the server and the mobile device, sending the data request and receiving the data. When the time slot has longer duration, the system will not be able to adapt with fast channel variations, and take advantage of better strategies and transmission decisions, which affects the quality perceived at the user end. Therefore, a large time slot duration may lead to only one decision if the video size is small. As shown in Figures 4.14, 4.15 and 4.16, the performance decreased with the increase of the time slot duration. Our proposed approach TS-PQ decides every time slot duration on the transmission strategy based on the WiFi and cellular rates

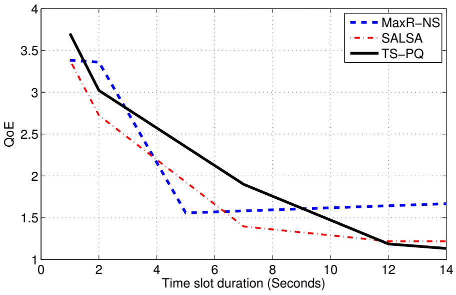


Figure 4.16: QoE MOS variation with respect to time slot duration T_s.

estimation. The accuracy of the rates estimation decreases when the time slot duration is longer. This results in inaccurate and inefficient transmission decisions affecting the overall performance of the approach. Accordingly, the solution needs to be real-time, adapt to the fast variation of the channel conditions to provide accurate rate estimation; for these reasons, we chose to use 1 second time slot duration in our work.

4.5 Test Bed Implementation for Cellular/WiFi Traffic Splitting under Realistic Operational Conditions

In the previous section, the proposed real-time QoE-aware resource management approach for video streaming applications was evaluated using simulations. In this section, the approach is tested under realistic operational conditions for accurate evaluation and validation using our own test bed implementation.

4.5.1 Test Bed Setup

In our test bed implementation, we considered the different components of the HetNet architecture; they can be represented by the following three levels: (i) the application service provider layer, (ii) the network wireless interfaces and operator level, and (iii) the user end as shown in Figure 4.17. The test bed is implemented using a modular approach which facilitated enhancements and extensions to implement and test various protocols, design alternatives, or intelligence options.

- **The application service provider layer:** Video streaming applications are deployed on an HTTP server acting as application service provider, and the source for the video files. The server receives data requests from the

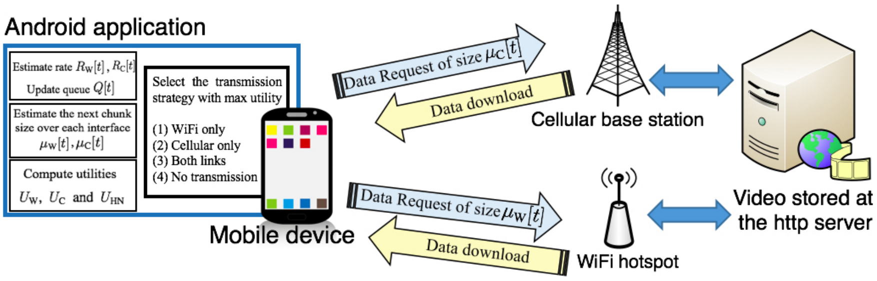

Android application
Estimate rate $R_{\mathrm{W}}[t]$, $R_{\mathrm{C}}[t]$
Update queue $Q[t]$
Estimate the next chunk size over each interface $\mu_{\mathrm{W}}[t]$, $\mu_{\mathrm{C}}[t]$
Compute utilities U_{W}, U_{C} and U_{HN}
Select the transmission strategy with max utility
(1) WiFi only
(2) Cellular only
(3) Both links
(4) No transmission
Mobile device
Data Request of size $\mu_{\mathrm{C}}[t]$
Data download
Cellular base station
Data Request of size $\mu_{\mathrm{W}}[t]$
Data download
WiFi hotspot
Video stored at the http server

Figure 4.17: Test bed implementation composed of three components:(1) mobile device application, (2) network interfaces, and (3) HTTP application server.

mobile device over WiFi and cellular networks, with specific data size and offset to be downloaded.

- **The network wireless interfaces and operator level:** In our conducted experiments, we used two wireless interfaces WiFi (802.11b) and 3G cellular networks.

- **The user end:** The client application is implemented using Java programming language on an Android mobile device. The decision on the amount of data to be downloaded over each interface every time slot is made, as decribed in section 7.2, based on the one of the following selected transmission strategy: (1) WiFi only, (2) cellular only, (3) both links simultaneously, or (4) no transmission. Accordingly, a specific amount of data is then requested at time slot t from the server using cellular and WiFi links, respectively. The requests are sent to the HTTP server indicating the offset and the size of the requested data. The data downloaded is reassembled as frames and played on the device.

The main challenge in the test bed implementation is to allow the use of multiple interfaces simultaneously, in addition to implementing the proposed algorithms and test them under real-time conditions. In the current mobile devices, the traffic is offloaded directly to WiFi when WLAN is available. Some new smartphones such as iOS 9 Iphones, Samsung Galaxy S5 and Sony Experia Z3 introduced auto-switching between WiFi and cellular data networks to avoid poor WiFi connections. However, traffic splitting and the use of multiple networks simultaneously is not yet supported. Our test bed design addresses this issue by supporting both techniques and giving the opportunity to the device to be connected to the best network for data download or using both links simultaneously to achieve performance gains. This can be achieved by allowing parallel transmission using the concept of multi-threading on a rooted Android device.

The user downloads a video with specific size, duration and frame rate. For real-time decision making, at each time slot of duration 1 second, the client application makes decision on the links to use for downloading data based on the selected strategy. If the transmission rate is lower than the video arrival rate, the mobile device downloads only a fraction of the requested video data. The remaining data that was not downloaded is stored in the queue to be downloaded in the next time slots. The video data is not lost, the frames are not skipped, they are only delayed when stalls exist. If the transmission rate is higher than the video arrival rate, the data is downloaded on time without any delay, stalls or freeze frames. In this case, the queue is empty with a queue size of zero. To compare the performance of the various strategies mentioned above, we based our evaluation on performance metrics such as the queue size, the average throughput, total energy consumption, delay and QoE.

Since the rates are varying and dependent on the channel conditions, the comparison of the algorithms may not be accurate. For these reasons, we developed an emulator for transmission rate control. The emulator controls the transmission rate and provides the same rate distribution every time for fair comparison between algorithms. To represent realistic scenarios, the WiFi and cellular networks were monitored, the values for WiFi and cellular transmission rates were recorded as traces. These traces were fed into the emulator; which restricts the downloading rates to the values in the traces. Accordingly, the performance of the algorithms was tested under same conditions for accurate evaluation and comparison.

4.5.2 Test Bed Experimental Results

In our considered scenario, the video has a size of 7 MBytes, duration of 60 seconds, and frame rate of 25 fps. The arrival rate will be 117 Kbytes every second. The WiFi and cellular transmission rates traces collected in Beirut Lebanon, R_{W} and R_{C}, presented an average of 42 KBps and 46 KBps, respectively. Using the same analysis presented in Section 4.4.4, $V_1[t]$ and $V_2[t]$ were chosen to be $1.25 \cdot 10^{11}$ and $3 \cdot 10^{10}$, respectively. The results for the different approaches are presented in Table 4.5. In addition, we considered different scenario with R_{W} and R_{C} average rates of 130 KBps. $V_1[t]$ and $V_2[t]$ chosen to be $2.5 \cdot 10^{10}$ and $9 \cdot 10^{12}$, respectively. The results are shown in Table 4.6.

Similar to the simulations analysis, the results showed that our proposed approach was able to provide the best balance between QoE, delay and energy consumption. It provided high user satisfaction with an acceptable increase in energy consumption. Note that the number of stalls is high in Table 4.5 since the WiFi and cellular rates were relatively low in the traces, which led to a poor QoE for all the compared algorithms.

The test bed results presented in Table 4.6 are highly correlated to the simulations results presented in Tables 4.1 and 4.3. Similar analysis presented in Section 4.4.4 can be obtained when comparing the implemented approaches. The results show low performance for WiFi only, cellular only and network selection

Table 4.5: Test bed results and statistics with R_{W} and R_{C} average of 42 KBps and 46 KBps, respectively

	WO	**CO**	**MaxR-NS**	**MinE-NS**	**SALSA**	**TS-S**	**TS-P**	**TS-PQ**
Total streaming time (s)	169.72	154.96	144.31	169.11	170.11	72.76	117.33	73.26
Total delay (s)	109.72	94.92	84.31	109.11	110.11	12.76	57.33	13.26
Average throughput (KBps)	42.21	46.24	49.65	42.37	42.22	100.73	61.06	98.45
Total energy consumption (J)	221.82	287.29	214.14	211.02	222.32	224.83	218.51	219.24
Average queue size (KB)	2374.21	2444.66	2367.91	2294.07	2317.01	858.06	1212.97	1106.96
Maximum queue size (KB)	4626.59	4554.67	4569.81	4567.51	4588.73	1524.61	2270.57	7165.45
Number of stalls	173	175	165	174	175	52	99	50
QoE ϕ (4.15)	1	1	1	1	1	1	1	1

Table 4.6: Test bed results and statistics with R_{W} and R_{C} average of 130 KBps

	WO	**CO**	**MaxR-NS**	**MinE-NS**	**SALSA**	**TS-S**	**TS-P**	**TS-PQ**
Total streaming time (s)	62.02	61.87	61.81	61.81	61.97	60	60.58	60
Total delay (s)	2.02	1.87	1.81	1.81	1.97	0	0.58	0
Average throughput (KBps)	115.53	115.82	115.92	115.92	115.62	119.42	118.27	119.42
Total energy consumption (J)	79.72	112.63	112.03	79.27	80.31	95.92	81.08	84.51
Average queue size (KB)	152.14	137.60	132.00	125.37	147.84	0	12.96	0
Maximum queue size (KB)	241.66	223.08	216.58	215.68	235.84	0	69.82	0
Number of stalls	11	9	8	10	11	0	4	0
QoE ϕ (4.15)	1.80	2.25	2.51	2.01	1.80	5	3.74	5

Table 4.7: Test bed results and statistics with R_{W} and R_{C} average of 110 KBps and 450 KBps

	R_{W} and R_{C} average 110 KBps				**R_{W} and R_{C} average 450 KBps**			
	WO	**CO**	**MaxR-NS**	**TS-PQ**	**WO**	**CO**	**MaxR-NS**	**TS-PQ**
Total streaming time (s)	72.57	68.60	66.92	62.85	62.49	61.89	60.71	60
Total delay (s)	12.57	8.60	6.92	2.85	2.49	1.89	0.71	0
Average throughput (KBps)	98.7	104.4	107.1	114.0	114.66	115.78	118.03	119.42
Total energy consumption (J)	94.85	125.35	101.55	92.61	75.48	114.03	109.56	72.65
Average queue size (KB)	197.41	241.15	148.78	23.39	86.62	8.24	2.84	0
Maximum queue size (KB)	817.61	851.00	515.83	300.04	297.44	145.79	68.49	0
Number of stalls	22	22	15	6	6	3	2	0
QoE ϕ (4.15)	1.02	1.02	1.22	3.05	3.06	4.02	4.39	5.00

strategies. The user experienced more delay, freeze frames and stalls when one link is selected. Comparing the approaches considering traffic splitting, TS-P provided the lower energy consumption with a tradeoff cost in terms of QoE. The proposed TS-PQ approach provided a very high quality of experience similar to TS-S when both links are used simultaneously with 11.8% lower energy consumption. This proves the effectiveness of the proposed approach and demonstrates the feasibility of achieving performance gains in practice using standard mobile devices.

4.5.3 Validation: Test Bed versus Simulation Results

To validate our simulation results, we conducted scenarios where data rates are similar to those used in the simulation results (Sections 4.4.4). Table 4.7 presents test bed results for WiFi and cellular average rates of 110 KBps and 450 Kbps using the following approaches: (1) WiFi only (WO), (2) Cellular only (CO),

(3) Maximum rate network selection (MaxR-NS), and (4) our proposed traffic splitting approach with delay-power-QoE balance (TS-PQ).

The obtained test bed results are similar to the simulation results presented in Table 4.1 and Table 4.3. For instance, the simulated results for TS-PQ approach showed a total streaming time of 63.2 seconds with 7 stalls and a QoE of 2.74 when the average rate is 110 KBps (see Table 1). The test bed results for the same scenario showed a total streaming time of 62.85 seconds with 6 stalls and a QoE of 3.04 (see Table 4.7). TS-PQ led to high performance gains in both simulations and test bed results for average rates of 450 KBps. The total streaming time is 60 seconds without any stalls providing an excellent QoE (see Tables 4.1 and 4.7). The results presented in Table 4.7 are also coherent with the test bed results presented in Tables 4.5 and 4.6. Our proposed approach is shown to achieve enhanced performance and a balance between delay, energy consumption and QoE.

4.6 Dynamic Traffic Splitting Limitations and Challenges

This chapter provided a solution for real-time traffic splitting across cellular and WiFi heterogeneous networks that provides improved QoE while reducing energy consumption and delay. The solution is based on a Lyapunov drift-plus-penalty formulation. The performance of the proposed approach was evaluated using both simulations and our own test bed implementation under realistic operational conditions using video on demand streaming applications. Results for various scenarios demonstrated favorable performance for the proposed traffic splitting approach.

The proposed solution can be extended to handle other types of applications that may exhibit different experience for the user such as live video or large file downloads. In these cases, the QoE metrics need to be adjusted for the specific experience such as data loss or total delays. Additionally, it can be extended towards a more optimized solution by predicting downstream performance in future time slots, capturing the need and cost associated with data re-transmission to recover from losses, or accounting for video multi-casting in multi-user scenarios taking into account resource limitations.

Chapter 5

Traffic Offloading for Maximum User Capacity in Dense D2D Cooperative Networks

In Chapters 3 and 4, we addressed user-centric cellular/WiFi resource management strategies for a single user considering network selection and traffic splitting. In this chapter, we extend our work to consider multi-user cellular/WiFi resource management in ultra dense device-to-device cooperative networks.

Ultra dense networks and device-to-device communications are expected to play a major role in 5G networks to meet tremendous traffic requirements. Typical UDNs include dense urban areas, open-air assembly and stadiums where a very high number of users request simultaneously large amounts of data [8]. In conventional networks, the mobile terminals receive their data from the base station or access point. In device-to-device cooperative networks, the mobile terminals take advantage of the co-existence of other mobile terminal to download the common content data. D2D cooperation is considered one of the main solution to increase the system capacity when macro network resources become scarce due to large traffic demands [9]. In addition, data caching and content owners distribution in D2D networks are considered promising ways to increase the system capacity and coverage. The content owners mobile terminals will distribute the common content data to other mobile terminals. Accordingly, the MTs can cooperate to receive data using different wireless interfaces, either from the BS/AP over long range links or from other mobile terminals or content owners over short range D2D links.

One of the major challenges in UDNs is the limited number of non-overlapping orthogonal channels. The use of non-orthogonal channels causes interference which decreases the achievable user throughput. Accordingly, considering channel allocation along with traffic offloading is needed to ensure that the users are provided with their service target rate.

In our work, we formulated the resource management problem where the channel allocation problem is simultaneously considered with traffic offloading taking into account the interference caused by the use of non-orthogonal channels in our problem formulation. The problem can be shown to be NP-Complete and cannot be solvable for a large number of users. Due to the high complexity of the problem, we divided the resource management problem into two optimization problems. First, we address optimal traffic offloading maximizing use capacity, where the channels are assumed to be orthogonal. The solution determines the best strategy for downloading the content either over long range connectivity from the access points or short range connectivity from peer mobile devices. We then solve optimal channel allocation maximizing the distance between co-channels transmitters, thus, reducing interference.

In this chapter, we address traffic offloading in dense D2D cooperative network where a large number of users request simultaneously common content data with/without content owners caching. We formulate the problem as an optimization problem to find the optimal LR and SR channel allocation constrained by the number of APs, LR and SR channels, users per cooperation cluster, and transmission rate. We solve the optimization problem using AIMMS software and CPLEX as a solver. To evaluate our solution, we focus on a stadium topology to demonstrate the significant gains of optimized traffic offloading in ultra dense wireless networks.

Optimal solutions may not be achievable in real-time ultra dense D2D cooperative networks due to the high time complexity of the problem. In addition, the optimization problem is holistic and considers all the existing users when providing optimal traffic offloading solutions. Accordingly, with the arrival of new users, the channel assignments and connections will change to find optimal LR and SR channel allocation. This leads to frequent changes in the allocation of channels and the role of users as cluster heads, LR and SR users, which is not feasible in real-time networks. For these reasons, we propose a dynamic tree-based traffic offloading approach (TBTO) which assigns the users' connections consecutively based on a tree having BSs/APs and mobile terminals as nodes. We show that the proposed approach is able to provide near-optimal solutions with very low time complexity.

This chapter is organized as follows. The system model is presented in Section 5.1. The resource management optimization problem formulation considering simultaneously traffic offloading and channel allocation is detailed in Section 5.2. The optimal traffic offloading for maximum user capacity in dense D2D cooperative networks is presented in Section 5.3. The proposed dynamic tree-based traffic offloading approach is presented in Section 5.4. Performance results are presented in Section 5.5. Limitations and challenges are drawn in Section 5.6.

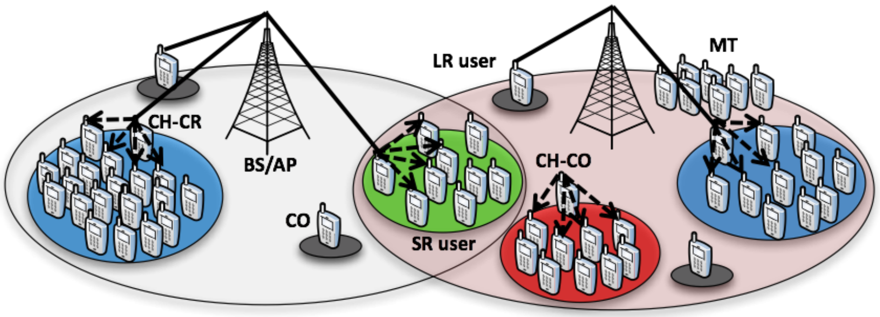


Figure 5.1: D2D cooperative network formed by one BS/AP and four clusters served by three CH-CRs and one CH-CO. CH-CRs receives over LR from BS/AP and transmit to SR users. Three COs are present: one CO serves as CH-CO transmits to other SR users and two considered served.

5.1 System Model

In our work, we address resource management in cache-enabled D2D cooperative networks where some users are downloading common content video streaming data while others have the data cached. The mobile terminals can use two wireless interfaces: one interface to communicate with the BS/AP over a long range wireless technology (such as WLAN, UMTS/HSPA, or LTE) and another interface to communicate with other MTs or content owners using a short range wireless technology (such as LTE-Direct, WiFi-Direct, Bluetooth or WiFi ad hoc mode).

As shown in Figure 5.1, the network is formed by BSs/APs, content owners and clusters served by a cluster head mobile terminal. Accordingly, a mobile terminal may receive its data from a BS/AP over LR wireless technology or from another MT using SR wireless technology. In our work, we define: (1) LR user as a mobile terminal receiving data from a BS/AP over LR connection, (2) SR user as a mobile terminal receiving data from other MT acting as a cluster head (CH) over SR connection, (3) a CH is a mobile terminal transmitting data to SR users over SR channels. A cluster head can be either:(1) a content owner (CH-CO) which is a mobile terminal who has already the data content cached, or (2) a content recipient (CH-CR) who is a LR user receiving data from BS/AP over LR channel.

Our network is formed by M BSs/APs and a large number of mobile terminals K. A MT i requests a common content data such as live and on-demand video streaming with a specific transmission target rate $R_{\mathrm{T},i}$. Each small cluster composed of mobile terminals served by the same mobile terminal CH is considered a cluster. Accordingly, one cluster head exists per cluster or group. A content owner has already the data cached, therefore, we consider a CO i served without receiving the data over LR or SR connections. A CO can act as a cluster head CH-CO when it transmits data to other MTs. In our formulation, we used input parameter c_i to indicate whether a mobile terminal i is a content owner.

We assume that the rate adaptation is based on M-QAM modulation. The rates $R_{\mathrm{L},ij}$ and $R_{\mathrm{S},ij}$ are the rates achievable on LR and SR channels between transmitter i and receiver j, respectively, computed as $R_{ij} = \log_2 M_{ij} \cdot W$, where W is the passband bandwidth of the channel and assuming the symbol rate is equal to the passband bandwidth, and M_{ij} is the highest possible order of the M-QAM modulation scheme selected based on the following expression [114]:

$$P_{\mathrm{e}} \leq 0.2e^{-1.5\gamma_{ij}/(M_{ij}-1)} \tag{5.1}$$

where P_{e} is the target probability of error, and γ_{kj} represents the signal-to-noise ratio (SNR) when the LR and SR channels are orthogonal. It represents the signal-to-interference-plus-noise ratio (SINR) when the channels are non-orthogonal.

In case the LR and SR channels are orthogonal, the signals will not interfere each other. γ_{kj} will be the signal-to-noise ratio given by (5.2):

$$\gamma_{ij} = P_{r,ij}/\sigma^2 \tag{5.2}$$

where σ^2 is the thermal noise power and $P_{\mathrm{r},ij}$ is the received power linked to the transmit power $P_{\mathrm{t},ij}$ of the transmitter i as follows:

$$\left(\frac{P_{\mathrm{r},ij}}{P_{\mathrm{t},ij}}\right)_{dB} = 10\log_{10}\kappa - 10\alpha\log_{10}\frac{d_{ij}}{d_0} + (h_{ij})_{dB} \tag{5.3}$$

where κ is a pathloss constant which depends on the antenna characteristics and wireless environment, α is the pathloss exponent, d_0 is a reference distance (typically 1 or 10 meters in indoor or short range outdoor scenarios), d_{ij} is the distance between transmitter i and receiver j, and h_{ij} is a random variable representing channel fading [114].

In case the channels are non-orthogonal, interference is seen by the mobile terminals. In general, a MT will be subjected to interference from different BSs/APs using non-orthogonal channels while receiving data over LR, and from cluster heads using non-orthogonal channels while receiving over SR. In our case model, we consider a network composed by a limited number of base stations. We then assume that the LR channels allocated are orthogonal. Therefore, the LR mobile terminals will not be subjected to LR interference. However, the SR users will be subjected to interference caused by cluster heads using non-orthogonal channels. The SINR is represented by γ_{kj} as follows:

$$\gamma_{ij} = \frac{P_{r,ij}}{\sigma^2 + \sum_{c\in\mathscr{C}_i} P_{r,c}} \tag{5.4}$$

where $P_{r,ij}$ is the received power from cluster head i over SR, $P_{r,cj}$ is the received power as interference, and c are cluster heads that belong to $\mathscr{C}_i$, defined as set of cluster heads transmitting on same channel as cluster head i.

Accordingly, based on the transmit power, distance and channel conditions between the receiver and the transmitter, the transmission rate is estimated. The goal is to serve the maximum number of users, while meeting the service target rate for all the served MTs subject to system constraints and network bandwidth limitations.

5.2 Resource Management Optimization Problem Formulation with Caching Considerations

In this section, we formulate the resource management problem as an optimization problem while considering traffic offloading and channel allocation simultaneously. The goal is to maximize the capacity of the system by offloading the data traffic to D2D communication taking advantage of the existence of content owners caching. We present first the optimization problem formulation including objective function, decision variables and constraints. We then present solution methodology and complexity analysis.

5.2.1 Resource Management Problem Formulation

We aim at serving the maximum number of mobile terminals with minimum number of LR channels while maintaining system target performance for every user. The system takes into consideration the existence of content owners which are mobile terminals having the data cached. The problem is formulated as an optimization problem aiming at determining the download strategy for every user while meeting target transmission rate. The outcome of the solution determines the mobile terminal connectivity for downloading data either from BS/AP via long range connectivity or from another MT via short range connectivity, in addition to the allocation of channels to cluster heads.

The decision variables (Table 5.1) are presented as follows:

- z_i: a binary variable that indicates whether mobile terminal i is served, i.e., receiving data via LR from a BS/AP or via SR from another mobile terminal. In general, users might not be served due to capacity and/or coverage limitation.

$$z_i = \begin{cases} 1 & \text{if MT } i \text{ is served} \\ 0 & \text{otherwise} \end{cases} \tag{5.5}$$

- y_{mi}: a binary variable that indicates whether mobile terminal i is receiving data over LR from BS/AP m.

$$y_{mi} = \begin{cases} 1 & \text{if MT } i \text{ receives data from BS/AP } m \\ 0 & \text{otherwise} \end{cases} \tag{5.6}$$

- v_{ij}: a binary variable that indicates whether mobile terminal j is receiving data over SR from MT i.

$$v_{ij} = \begin{cases} 1 & \text{if MT } i \text{ transmits data to MT } j \\ 0 & \text{otherwise} \end{cases} \tag{5.7}$$

- u_i: a binary variable that indicates whether mobile terminal i is a cluster head, i.e., receiving data over LR and transmitting over SR to other mobile terminals.

$$u_i = \begin{cases} 1 & \text{if MT } i \text{ is a cluster head} \\ 0 & \text{otherwise} \end{cases} \tag{5.8}$$

- Q_{pi}: binary variable that indicates whether channel p is allocated to mobile terminal i.

$$Q_{pi} = \begin{cases} 1 & \text{if channel } p \text{ is allocated to cluster head } i \\ 0 & \text{otherwise} \end{cases} \tag{5.9}$$

Accordingly, the decision variables are: $\mathbf{y}$ a matrix of size $M \times K$, $\mathbf{v}$ a matrix of size $K \times K$, $\mathbf{Q}$ a matrix of size $K \times N_{\text{SRo}}$ $\mathbf{u}$ and $\mathbf{z}$ vectors of length K. The resource management problem is subjected to several constraints in terms of capacity and coverage limitations. The problem is formulated as follows:

$$\underset{\mathbf{y},\mathbf{v},\mathbf{u},\mathbf{z},\mathbf{Q}}{\text{argmin}} \quad \sum_{m=1}^{M}\sum_{i=1}^{K} y_{mi} - \beta \sum_{i=1}^{K} z_i \tag{5.10}$$

subject to

$$v_{ij} \leq \sum_{m=1}^{M} y_{mi} + c_i, \forall i, \forall j \tag{5.11}$$

$$c_j + \sum_{i=1, i\neq j}^{K} v_{ij} + \sum_{m=1}^{M} y_{mj} = z_j, \forall j \tag{5.12}$$

$$\sum_{m=1}^{M}\sum_{i=1}^{K} y_{mi} \leq N_{\mathrm{LR}} \tag{5.13}$$

$$\sum_{i=1}^{K} y_{mi} \leq K_{\mathrm{L},m}, \forall m \tag{5.14}$$

$$\sum_{j=1, j\neq i}^{K} v_{ij} \leq K_{\mathrm{C},i}, \forall i \tag{5.15}$$

$$u_i \leq \sum_{m=1}^{M} y_{mi} + c_i, \forall i \tag{5.16}$$

$$u_i \geq v_{ij}, \forall i, \forall j \tag{5.17}$$

$$\sum_{i=1}^{K} u_i \leq N_{\mathrm{SR}} \tag{5.18}$$

$$\sum_{m=1}^{M} R_{\mathrm{L},mi} \cdot y_{mi} + \sum_{j=1, j\neq i}^{K} R_{\mathrm{S},ji} \cdot v_{ji} \geq R_{\mathrm{T},i} \cdot z_i \cdot (1 - c_i), \forall i \tag{5.19}$$

$$\sum_{p=1}^{N_{\mathrm{SRo}}} Q_{pi} = u_i, \forall i \tag{5.20}$$

$$\left\lfloor \frac{\sum_{i=1}^{K} u_i}{N_{\mathrm{SRo}}} \right\rfloor \leq \sum_{i=1}^{K} Q_{pi} \cdot u_i \leq \left\lceil \frac{\sum_{i=1}^{K} u_i}{N_{\mathrm{SRo}}} \right\rceil, \forall p \tag{5.21}$$

$$y_{mi} \in \{0,1\}, v_{ij} \in \{0,1\}, u_i \in \{0,1\}, z_i \in \{0,1\}, Q_{pi} \in \{0,1\} \tag{5.22}$$

- Equation (5.10) is the objective function which aims to miminize the usage of long range channels and maximize coverage and thus force more cooperation between mobile terminals. The aim is to serve the largest number of users while using the minimum LR channels. β is a positive coefficient indicating the impact of maximizing the number users served. Since our primary goal is to serve the maximum number of users, β parameter should have a high value. We assume β is equal to the number of active users to give very high impact for serving users instead of minimizing the use of LR channels.

- The first constraint (5.11) guarantees that only a mobile terminal i receiving over LR (CH-CR) or a content owner can forward data to other mobile terminals over SR. For instance, MT j can receive data from MT i ($v_{ij} = 1$) only if MT i receive data over LR from a BS/AP m $\left(\sum_{m=1}^{M} y_{mi} = 1\right)$ or if MT i is a content owner with $c_i = 1$. Otherwise, if MT i is not a CO and does not receive over LR $\left(\sum_{m=1}^{M} y_{mi} = 0\right)$, MT j cannot receive data from MT i ($v_{ij} = 0$).

- The second constraint (5.12) makes sure any MT j that is served by the system ($z_j = 1$), receives data either on LR from BS/AP m $\left(\sum_{m=1}^{M} y_{mj} = 1\right)$ or SR from mobile terminal i (one of v_{ij} variables is equal to 1) or is a content owner $c_j = 1$.

- Constraint (5.13) ensures that the number of LR users is less than N_{LR}.

- Constraint (5.14) guarantees that the number of users served by a BS/AP m is less than $K_{\text{L},m}$.

- Constraint (5.15) guarantees that the number of users in a cluster served by a cluster head i is less than the maximum allowed number $K_{\text{C},i}$.

- Constraints (5.16), (5.17) and (5.18) guarantee that the number of clusters is less than N_{SR}. The variable u_i indicates if MT i is a cluster head which can be a cluster head content recipient CH-CR or a content owner CH-CO. Constraint (5.16) guarantees that MT i can be a cluster head ($u_i = 1$) if it is a content owner $c_i = 1$ or it receives over LR (one of the y_{mi} variable is equal to 1), constraint (5.17) ensures that MT i can be a cluster head if it transmits data over SR (one of the v_{ij} variable is equal to 1). Constraint (5.18) limits the number of cluster heads to N_{SR}.

- Constraint (5.19) ensures that the throughput for every mobile terminal (considered served and is not a CO, i.e., $z_i = 1$ and $c_i = 0$) is greater than target rate $R_{\text{T},i}$. If a MT is receiving data over LR, its rate R_i will be equal to $R_{\text{L},mi}$ with one of y_{mi} variable equals to 1 and $v_{ji} = 0, \forall j$. If the mobile terminal i is receiving over SR from another MT j, R_i will be equal to $R_{\text{S},ji}$ with $y_{mi} = 0, \forall m$ and $v_{ji} = 1$.

- Constraint (5.20) ensures that every cluster head mobile terminal is assigned one SR channel.

Table 5.1: Main parameters and variables

Parameters	
K	the set of requesting MTs, where a MT is referred to as MT i, $i = 1, ..., K$
M	the set of BSs/APs, where a BS/AP is referred to as BS/AP m, $m = 1, ..., M$
β	positive coefficient indicating the tradeoff between maximizing the number of users served and minimizing the number of LR channels
c_i	a binary variable that indicates whether MT i is a content owner
N_{SRo}	number of orthogonal non-overlapping SR channels
d_{ij}	distance between transmitter (BS/AP or MT) i and MT j
$R_{\mathrm{S},ij}$	transmission rate on SR from the MT i to MT j
$R_{\mathrm{L},mj}$	transmission rate on LR from the BS/AP m to MT j
$R_{\mathrm{T},i}$	target transmission rate to MT i to meet video application requirements
N_{LR}	maximum number of LR channels in the network
N_{SR}	maximum number of clusters in the network
N_{CH}	number of cluster heads in the network
$K_{\mathrm{L},m}$	maximum number of MTs served by a BS/AP m
$K_{\mathrm{C},i}$	maximum number of MTs served over SR by a cluster head MT i in a cluster
Variables	
z_i	a binary variable that indicates whether MT i is receiving data
y_{mi}	a binary variable that indicates whether MT i is receiving data over LR from BS/AP m
v_{ij}	a binary variable that indicates whether MT j is receiving data over SR from MT i
u_i	a binary variable that indicates whether MT i is a cluster head
Q_{pi}	a binary variable that indicates whether channel p is allocated to cluster head i

- Constraints (5.21) ensure that all the channels are used with minimum reuse factor. If the number of orthogonal channels N_{SRo} is greater than the number of cluster heads N_{CH} ($\sum_{i=1}^{K} u_i$), every channel p can be then assigned maximum once $\left(0 \leq \sum_{i=1}^{K} Q_{pi} \cdot u_i \leq 1\right)$. If the number of orthogonal channels is less than the number of CHs, constraint (5.21) ensures that a channel p is then allocated with minimum reuse, i.e. maximum of $\left\lceil \frac{\sum_{i=1}^{K} u_i}{N_{\mathrm{SRo}}} \right\rceil$ and minimum of $\left\lfloor \frac{\sum_{i=1}^{K} u_i}{N_{\mathrm{SRo}}} \right\rfloor$.
- The last constraint sets the decision variables $\mathbf{y}$, $\mathbf{v}$, $\mathbf{u}$, $\mathbf{z}$ and $\mathbf{Q}$ to be binary.

The user throughput can then be computed as presented in Section 5.1. The SINR can be presented as follows:

$$\gamma_{ij} = \frac{P_{r,ij}}{\sigma^2 + \sum_{h=1,h\neq i}^{K} \sum_{p=1}^{N_{\mathrm{SRo}}} u_h \cdot Q_{ph} \cdot Q_{pi} \cdot P_{r,hj}} \tag{5.23}$$

where cluster head i sends data to mobile terminal j. The cluster heads h will cause interference to the main transmission between mobile terminal i and j.

5.2.2 Solution Methodology

The problem is a binary linear programming problem. The number of binary variables is $(M + N_{\mathrm{SRo}})K + K^2$ composed of: $\mathbf{y}$ a matrix of size $M \times K$, $\mathbf{v}$ a matrix of size $K \times K$, $\mathbf{Q}$ a matrix of size $K \times N_{\mathrm{SRo}}$, $\mathbf{u}$ and $\mathbf{z}$ vectors of length K that can be computed using values of variables $\mathbf{y}$ and $\mathbf{v}$. The problem is NP-complete. Starting with the first two constraints, the problem is divided into two directions or directed graphs whether the mobile terminal is receiving data over LR or SR. Each graph is composed of different subgraphs based on the problem constraints. The problem is similar to Minimum Dominating Set problem in Directed Graphs which has been shown as NP-Hard [115][116]. It can be shown that solution for the optimal resource management problem can be verified in polynomial time, thus the problem is NP-complete.

Due to its very high complexity, we divided the resource management problem into two separate problems: (1) optimal traffic offloading for maximum user capacity assuming all the channels are orthogonal (Section 5.3), and (2) optimal SR channel allocation to the cluster heads defined by the traffic offloading solution (Chapter 6).

5.3 Traffic Offloading Optimization Problem Formulation with Caching Considerations

In this section, we present the traffic offloading problem formulation including objective function, decision variables and constraints while assuming the channels are orthogonal. We then present solution methodology and complexity analysis.

Similar to the formulation presented in Section 5.2, the decision variables are:

- z_i: a binary variable that indicates whether mobile terminal i is served.
- y_{mi}: a binary variable that indicates whether mobile terminal i is receiving data over LR from BS/AP m.
- v_{ij}: a binary variable that indicates whether mobile terminal j is receiving data over SR from MT i.
- u_i: a binary variable that indicates whether mobile terminal i is a cluster head, i.e., receiving data over LR and transmitting over SR to other mobile terminals.

Accordingly, the decision variables are: $\mathbf{y}$ a matrix of size $M \times K$, $\mathbf{v}$ a matrix of size $K \times K$, $\mathbf{u}$ and $\mathbf{z}$ vectors of length K.

The traffic offloading problem can be formulated as follows:

$$\underset{\mathbf{y},\mathbf{v},\mathbf{u},\mathbf{z}}{\text{argmin}} \quad \sum_{m=1}^{M}\sum_{i=1}^{K} y_{mi} - \beta \sum_{i=1}^{K} z_i \tag{5.24}$$

subject to

$$v_{ij} \leq \sum_{m=1}^{M} y_{mi} + c_i, \forall i, \forall j \tag{5.25}$$

$$c_j + \sum_{i=1, i\neq j}^{K} v_{ij} + \sum_{m=1}^{M} y_{mj} = z_j, \forall j \tag{5.26}$$

$$\sum_{m=1}^{M}\sum_{i=1}^{K} y_{mi} \leq N_{\text{LR}} \tag{5.27}$$

$$\sum_{i=1}^{K} y_{mi} \leq K_{\text{L},m}, \forall m \tag{5.28}$$

$$\sum_{j=1, j\neq i}^{K} v_{ij} \leq K_{\text{C},i}, \forall i \tag{5.29}$$

$$u_i \leq \sum_{m=1}^{M} y_{mi} + c_i, \forall i \tag{5.30}$$

$$u_i \geq v_{ij}, \forall i, \forall j \tag{5.31}$$

$$\sum_{i=1}^{K} u_i \leq N_{\text{SR}} \tag{5.32}$$

$$\sum_{m=1}^{M} R_{\text{L},mi} \cdot y_{mi} + \sum_{j=1, j\neq i}^{K} R_{\text{S},ji} \cdot v_{ji} \geq R_{\text{T},i} \cdot z_i \cdot (1 - c_i), \forall i \tag{5.33}$$

$$y_{mi} \in \{0,1\}, v_{ij} \in \{0,1\}, u_i \in \{0,1\}, z_i \in \{0,1\} \tag{5.34}$$

The traffic offloading objective function and constraints (5.24) to (5.34) are exactly similar to the resource management objective function and constraints (5.10) to (5.22) presented in Section 5.2, respectively. However, the complexity of resource management optimization problem is reduced to consider traffic offloading problem without channel allocation. The decision variable $\mathbf{Q}$ indicating channel allocation is not considered in addition to constraints (5.20) and (5.21).

The problem is a binary linear programming problem. The number of binary variables is $MK + K^2$ composed of: $\mathbf{y}$ a matrix of size $M \times K$, $\mathbf{v}$ a matrix of size $K \times K$, $\mathbf{u}$ and $\mathbf{z}$ vectors of length K that can be computed using values of variables $\mathbf{y}$ and $\mathbf{v}$. It can be shown that solution for the optimal traffic offloading problem can be verified in polynomial time, thus the problem is NP-complete [115][116].

We solved the optimization problem using AIMMS software which is designed for modeling and solving large-scale optimization and scheduling-type problems. CPLEX optimization software package in AIMMS was used as a solver [117]. CPLEX uses the simplex algorithm to solve very large linear programming problems, convex and non-convex quadratic programming problems, and convex quadratically constrained problems.

5.4 Proposed Dynamic Tree-Based Traffic Offloading Algorithm

Due to the high complexity of the problem, optimal solutions may not be achievable in real-network dense scenarios. We propose dynamic tree-based traffic offloading TBTO algorithm to provide near-optimal solutions with very low time complexity. The proposed dynamic tree-based traffic offloading algorithm assigns the users' connections sequentially based on a tree having BSs/APs and mobile terminals as nodes. Our approach is based on a 4-level tree as follows: (1) the network as root node, (2) the BSs/APs as first level parent nodes, (3) the cluster heads and LR users as second level nodes receiving data from BSs/APs, and (4) SR users receiving data from cluster heads as fourth level terminal nodes.

In general, a mobile terminal can be connected to a BS/AP or another terminal considered as cluster head to download common content data. Accordingly, we consider BSs/APs or CHs as parent nodes. When connected to a BS/AP, a MT j is considered a LR user and is added to the tree. When connected to another mobile terminal i, MT j is added to the tree in the fourth level as a SR user. In addition, the mobile terminal i becomes a cluster head. Accordingly, every BS/AP forms its own sub-tree having LR users and cluster heads as their child nodes. Similarly, the CHs form their own sub-trees, called clusters in our model, having SR users as child nodes. We define $\mathcal{N}$ as the set of existing nodes in the tree composed of BSs/APs and assigned users.

Figure 5.2 presents a tree representation of a D2D network and the connections between the nodes. The network is formed by two BSs/APs and seven mobile terminals. Each BS/AP has its own sub-tree serving one CH and one LR user: BS/AP1 serves MT1 and MT4 and BS/AP2 serves MT5 and MT7. The cluster heads MT1 and MT5 form their own cluster where MT1 serves SR users MT2 and MT3, and MT5 serves SR user MT6.

To assign users' connections, the algorithm builds a tree considering users sequentially, one user at a time. It starts from the root node which is in our case the network. The root has by default all the BSs/APs as child nodes. To assign the connection for user i, the nodes $\mathcal{N}$ of the current tree are considered, and the mobile terminal i is added as a child node to the tree based on Algorithm 5.1.

In our work, we aim at maximizing network capacity by minimizing the num-

Algorithm 5.1: The proposed tree-based traffic offloading (TBTO) approach for maximum system capacity

```
Input:
- $K$          number of users,
- $M$          number of BSs/APs,
- $d_{ij}$     distance between transmitter (BS/AP or MT) $i$ and MT $j$,
- $N_{\mathrm{LR}}$   maximum number of LR channels in the network,
- $N_{\mathrm{SR}}$   maximum number of SR channels in the network,
- $K_{\mathrm{L},m}$  maximum number of MTs served by a BS/AP $m$,
- $K_{\mathrm{C},i}$  maximum number of MTs served by a cluster head MT $i$,
- $R_{\mathrm{T},i}$  target transmission rate to MT $i$ to meet video application requirements,

Output:
- $z_i$      a binary variable that indicates whether MT $i$ is receiving data
- $y_{mi}$   a binary variable that indicates whether MT $i$ is receiving data over LR from BS/AP $m$
- $v_{ij}$   a binary variable that indicates whether MT $j$ is receiving data over SR from MT $i$
- $u_i$      a binary variable that indicates whether MT $i$ is a cluster head

1: Consider the network as root node
2: Assign BSs/APs as default child nodes to the network root node. The BSs/APs
   will be then considered first level nodes in the tree. At this stage, $\mathcal{N} = \{m|m \in M\}$
3: Assign a connection and allocate a parent node for a MT $j$ as follows:
     1. Estimate the transmission rate $R_{\mathrm{X},nj}$ provided by a node $n$ existing in the
        tree $\mathcal{N}$, where $X$ represents either LR or SR connection:
          • Estimate transmission rate $R_{\mathrm{L},mj}$ over LR from a node $n$ which is
            in this case a BS/AP $m$
          • Estimate transmission rate $R_{\mathrm{S},ij}$ over SR from a node $n$ which can
            be a LR user or a cluster head MT $i$
     2. Consider the nodes providing transmission rates higher than the target
        rate of MT $j$ as candidate nodes CN.
        $CN = \{n|n \in \mathcal{N}, R_{\mathrm{X},nj} \geq R_{\mathrm{T},j}\}$
     3. Check if transmission rates satisfy MT $j$ target rate
          • if $CN = \emptyset$ then
          •   Delay MT $j$ assignment until all users are considered
          •   Add MT $j$ to the non-assigned users
          • else
          •   Consider the weight of SR rate $R_{\mathrm{S},nj}$ between the existing CH
            and LR users nodes $n$ and MT $j$ twice the weight of the LR rate
            $R_{\mathrm{L},nj}$ between the BSs/APs nodes $n$ and MT $j$ to encourage traffic
            offloading to D2D connectivity
```

```
3:    3.    • Check system constraints:
                  – if the candidate node $n$ is a CH MT $i$ & the number of its child nodes = $K_{\mathrm{C},i}$ then Eliminate node $n$ end if
                  – if the candidate node $n$ is a BS/AP $m$ & the number of its child nodes = $K_{\mathrm{L},m}$ then Eliminate node $n$ end if
                  – if the candidate node $n$ is a BS/AP $m$ & the number of LR users = $N_{\mathrm{LR}}$ then Eliminate node $n$ end if
                  – if the candidate node $n$ is a MT $i$ & the number of CHs = $N_{\mathrm{SR}}$ then Eliminate node $n$ end if
            • Select the candidate node $n$ providing higher system throughput
                  – if multiple nodes provide the same system throughput then Select node $n$ serving less child nodes end if
            • Add the MT $j$ as a child to node $n$
            • Update the tree, $\mathcal{N}$ and decision variables:
                  – if MT $j$ is added to the tree then MT $j$ is served & $z_j$ is set to 1 end if
                  – if node $n$ is a BS/AP $m$ then MT $j$ is a LR user & $y_{mj}$ is set to 1 end if
                  – if node $n$ is a MT $i$ then
                  –     MT $j$ is a SR user & $v_{ij}$ is set to 1
                  –     MT $i$ is a CH & $u_i$ is set to 1
                  – end if
            • end if
4: Repeat process (3) for all the non-assigned users until no more users can be added as nodes to the tree.
```

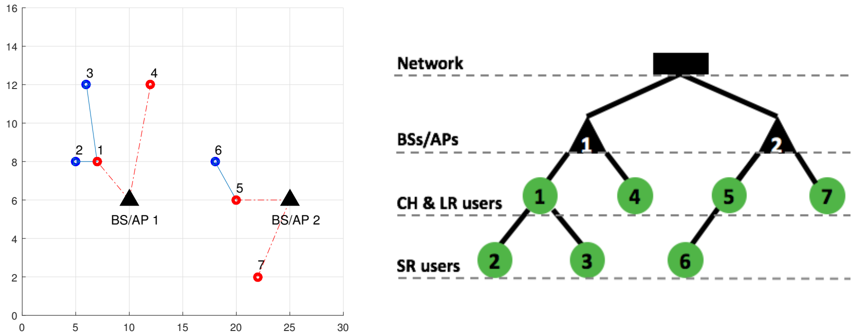


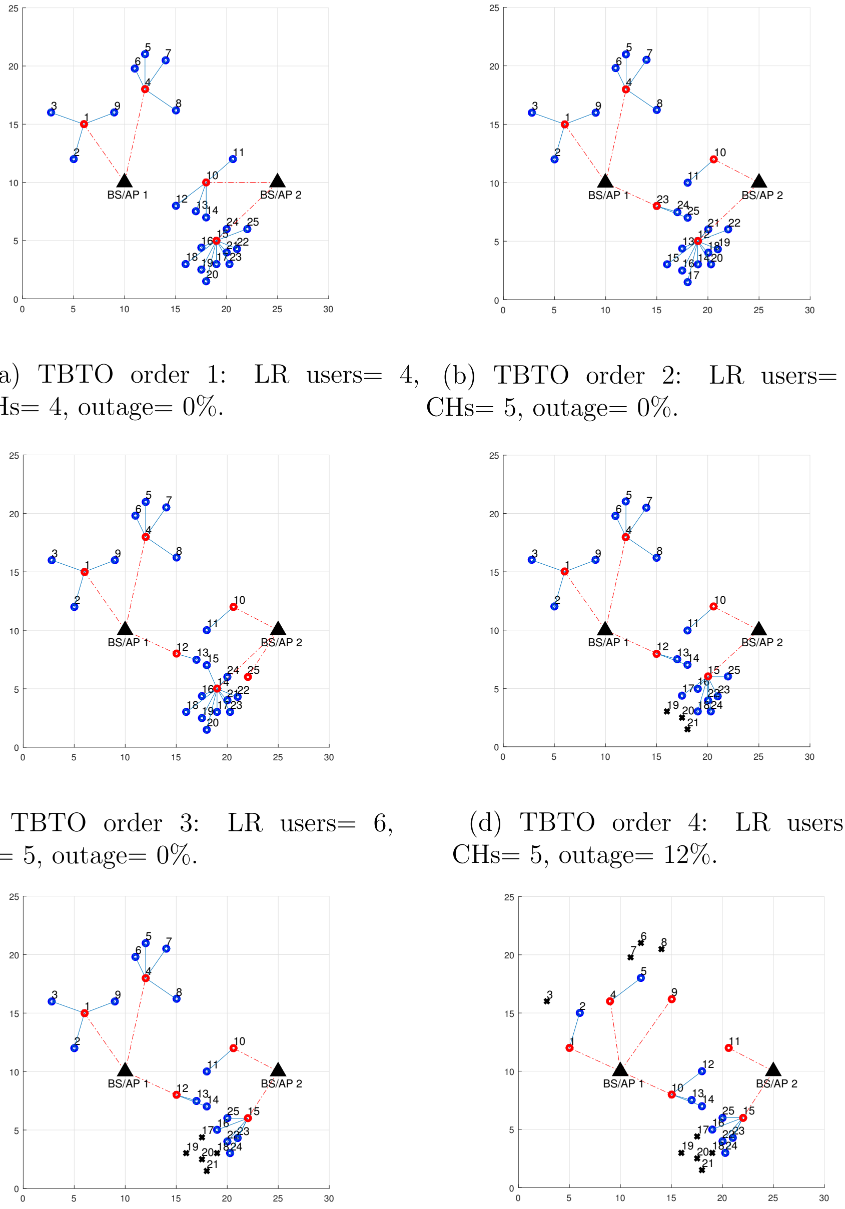
Figure 5.2: A tree representation of the network connections, formed by 4 levels: (1) network level, (2) BSs/APs level, (3) CHs and LR users level, and (4) SR users.

ber of LR channels used, which allows to serve more users. In order to increase the capacity of the network, the mobile terminals are encouraged to use D2D connectivity. Accordingly, we gave twice the weights of LR connectivity to D2D SR connectivity.

As presented in Algorithm 5.1, the connection for a mobile terminal j is assigned as follows: (1) the transmission rates provided by existing nodes $\mathcal{N}$: BSs/APs (level 2 nodes) and LR users and cluster heads (level 3 nodes) are estimated, (2) the nodes providing transmission rates higher than the target rate of MT j are considered as candidates nodes (CN), (3) the weight of SR rate $R_{\mathrm{S},nj}$ between the LR user n and MT j is considered twice the weight of the LR rate $R_{\mathrm{L},mj}$ between the BS/AP m and MT j to encourage D2D traffic offloading, and (4) select the candidate node providing higher system throughput.

In case two nodes provide the same throughput, the mobile terminal j will be connected to the node serving less MTs. The assignment also makes sure the system constraints are satisfied such as the maximum number of users in a cluster, the maximum number of LR users and SR users. If the target rate of a MT i is not satisfied by the nodes in the current tree, the mobile terminal assignment will be delayed and reconsidered later. This process continues until all users are considered and no more users can be added to the tree.

To illustrate an example of the proposed tree-based traffic offloading algorithm, we considered a network formed by two BSs/APs and 25 users. As presented in Figure 5.3(a), the proposed TBTO approach was able to provide an optimal solution where all the users are served with minimum number of LR channels and clusters. The network is divided into 4 clusters as follows: (1) MT 1 serving 3 MTs: 2, 3 and 9, (2) MT 4 serving 4 MTs: 5, 6, 7 and 8, (3)



(a) TBTO order 1: LR users= 4, CHs= 4, outage= 0%.

(b) TBTO order 2: LR users= 5, CHs= 5, outage= 0%.

(c) TBTO order 3: LR users= 6, CHs= 5, outage= 0%.

(d) TBTO order 4: LR users= 5, CHs= 5, outage= 12%.

(e) TBTO order 5: LR users= 5, CHs= 5, outage= 20%.

(f) TBTO order 6: LR users= 6, CHs= 4, outage= 36%.

Figure 5.3: Proposed tree-based traffic offloading approach performance for different user arrival orders. The network is formed by two BSs/APs and 25 users.

MT 10 serving 4 MTs: 11, 12, 13, and 14, and (4) MT 15 serving 10 MTs: 16 to 25. The solution of the proposed approach varies with the arrival order of users since the tree is built consecutively based on the user arrival. For instance, the order in Figure 5.3(b) shows sub-optimal assignment for the TBTO approach. The MT 11 in Figure 5.3(a) arrived before MT 10 and MTs 15 to 25 before MTs 12, 13 and 14. The TBTO solution in Figure 5.3(b) provides an outage 0% while using 5 LR channels serving as cluster heads. The order in Figure 5.3(c) provides an outage of 0% with 6 LR users where 5 serve as cluster heads. The cluster served by MT 14 reached its maximum number of users within a cluster, which made MT25 be served as LR user by BS/AP 2. Figure 5.3(d) and Figures 5.3(e) show outage of 12% and 20%, respectively, while forming 5 clusters. In Figure 5.3(d), MTs 19, 20 and 21 were out of coverage. In Figure 5.3(e), MTs 17, 18, 19, 20 and 21 were out of coverage. The worst performance provided by TBTO is presented in Figure 5.3(f) where the network is composed of 6 LR users, 4 cluster heads with an outage of 36%.

5.5 Traffic Offloading Results and Analysis

In this section, we present first the simulation setup including case study topology, assumptions and main system parameters. In Section 5.5.2, we present the performance of the network without D2D cooperation. In Section 5.5.3, we present solutions with D2D cooperation without content owner caching as a function of user density level. In Section 5.5.4, we consider content owner caching with D2D cooperation and show the system performance as a function of user density level varying the number of content owners. Performance results of the proposed dynamic tree-based traffic offloading algorithm are presented in Section 5.5.5.

5.5.1 Simulation Setup

As a case study, we consider a stadium with a capacity of 100,000 seats. Our topology is close to the topology of Camp nou stadium, located in Barcelona, Spain. Camp nou, with a seating capacity of 99354, is considered the largest stadium in Europe in terms of capacity. The considered dimensions of the stadium are assumed to be $280m \times 240m$ as shown in Figure 5.4. Due to the high complexity of the problem and the large number of users, we divided the area into small sections of $20m \times 20m$. The main system parameters are summarized in Table 5.2.

Mobile terminal demands

Every $20m \times 20m$ section is composed of 700 seats, assuming the seat width to be $0.5m$, and the space between rows to be $1.14m$ [118]. The attendants density

Table 5.2: Network parameters and assumptions

Parameters	Values
Section area	40m × 40m
Seats capacity C	2800 seats/section
Number of sections B	1, 2, or 4
Block area	40m × 40m for $B = 1$ 40m × 80m for $B = 2$
Number of BSs/APs per block	5 or 9
M	5×B or 9×B
A	0.1 - 1
K	$C \times B \times A$
$R_{\mathrm{T},i}$	1 Mbps
$K_{\mathrm{L},m}$ (K_{L})	30 connections/AP m
$K_{\mathrm{C},i}$ (K_{C})	10 connections/CH i
N_{LR}	$\sum_{m=1}^{M} K_{\mathrm{L},m}$
N_{SR}	$N_{\mathrm{LR}} \cdot K_{\mathrm{C}}$
P_{tLR}	10 Watts
P_{tSR}	0.5 Watts
W	0.5 MHz
P_{e}	10^{-3}
σ^2	10^{-3} Watts
κ	-31.54 dB
α	3.71
d_0	10 m

is then equal to 1.75 users per $1m^2$. We based our study on an area of $40m \times 40m$ composed of four $20m \times 20m$ sections, with total of 2800 seats and five deployed BSs/APs as presented in Figure 5.4. In our work, we refer to user activity A as the probability of users out of 2800, located in the target $40m \times 40m$ area and are simultaneously requesting common content distribution. The number of active mobile terminals K can then be computed as follows: $A \times 2800$. Non active users are not considered present in our model. The active mobile terminals are randomly distributed. In our considered scenarios, the MTs are assumed to download common content real-time video streaming with 1 Mbps target rate requirement.

Long range channels capacity and coverage

In our model, we assume the BSs/APs are using 2.4 GHz IEEE 802.11n WLAN. Assuming an overhead of 35%, interference of 35% and a maximum PHY rate of 72.2 Mbps, the estimated AP aggregated throughput will be around 30 Mbps. The maximum number of users served by one AP to meet the target requirements of 1 Mbps will be 30 users/AP. In our model, the transmit power of an AP is assumed to be 10 Watts.

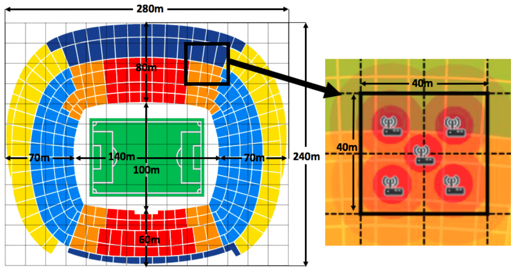


Figure 5.4: Stadium with dimensions $280m \times 240m$, divided into $20m \times 20m$ sections, each composed of 700 seats. An area of $40m \times 40m$ area is composed of four sections, 2800 seats and five deployed BSs/APs.

Table 5.3: Number of BSs/APs needed without D2D cooperation

Activity A	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1**
Active users in 40mx40m area	280	560	840	1120	1400	1680	1960	2240	2520	2800
BSs/APs needed in 40mx40m area	10	19	28	38	47	56	66	75	84	94
BSs/APs needed in 280mx240m area	360	684	1008	1368	1692	2016	2376	2700	3024	3384

Short range channels capacity and coverage

In our model, we assume the MTs are using 5GHz IEEE 802.11n WLAN. The maximum number of users K_{C} served by one cluster head is limited to a maximum of 10 users/cluster. The transmit power of a mobile terminal is assumed to be 0.5 Watts.

5.5.2 Performance Results: Conventional Model Without D2D Cooperation

In conventional networks, users download their data from BSs/APs using LR channels. In ultra dense networks, large number of users are requesting data simultaneously. Due to the limitation of capacity and bandwidth, a large number of BSs/APs is then required. Table 5.3 presents the number of BSs/APs required to serve simultaneously different user density and network activity. For a network with 0.1 low user activity, 10 BSs/APs are required to serve 280 users within a $40m \times 40m$ area. Our considered stadium is composed of 36 sections of $40m \times 40m$, accordingly, 360 BSs/APs are needed to serve low activity user density of 0.1 in a total $280m \times 240m$ area. The number of BSs/APs increases with the increase of the traffic demand and user activity due to the limitation of LR channels and users served per BS/AP. The number of APs reaches 94 to serve a high

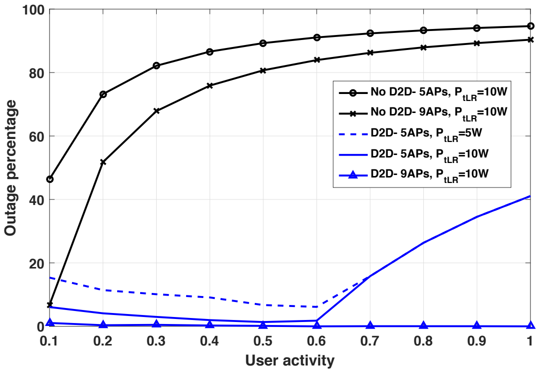


Figure 5.5: Outage percentage variation with user activity probability for different network scenarios composed of 5 or 9 APs with/without D2D cooperation.

density network composed of 2800 users. A total of 3384 BSs/APs are then needed to serve the high user density stadium. Deploying large number of APs is expensive and causes high interference due to the limit on the number of available IEEE802.11n orthogonal channels.

5.5.3 Performance Results: System Capacity Increase with D2D Cooperation

Figure 5.5 shows the outage probability for different scenarios with and without D2D cooperation while varying the user activity representing the density of the users requesting common content simultaneously within $40m \times 40m$ area. The outage probability is very high when no D2D cooperation is considered. The system capacity is limited to 150 and 270 possible connections over LR channels when 5 APs and 9 APs are deployed, respectively. This leads to an outage reaching 94.64% when the network density is very high.

The outage probability decreases when D2D cooperation is considered. The coverage range and capacity of the conventional network are extended by the cluster heads acting as providers to other MTs. When the network density increases within a specific area, the MTs are closer and tend to use SR channels for data download. Accordingly, the number of clusters formed increases to serve more users. The outage then decreases with the increase of the network density to reach a capacity limited by the number of LR channels K_{L} and the maximum number of users within a cluster K_{C} which is 1650 when 5 APs are deployed. For this reason, the outage probability increases for user activity more than 0.6 (1680 users). Increasing the transmit power $P_{\mathrm{t,LR}}$ of the APs from 5 Watts to

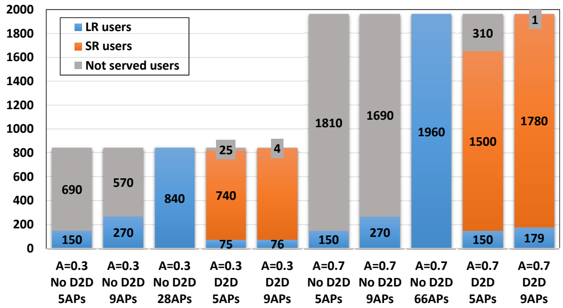


Figure 5.6: LR and SR channel allocation for low density (A=0.3) and high density (A=0.7) scenarios with/without cooperation.

10 Watts decreases the outage percentage from 15.35 to 6.07% for low user activity (A=0.1), and from 6.13 to 1.78% for high user activity (A=0.6). Deploying 9 APs allows the network to serve high user density. The outage probability is less than 1% when 2800 users are simultaneously requesting data.

Figure 5.6 shows the resource allocation of LR and SR channels for low density (A=0.3) and high density (A=0.7) network activity. The number of LR channels is very high when no D2D cooperation is used. The number of LR users is reduced from 840 (28 APs) when no D2D is considered to 75 (5 APs) and 76 (9 APs) with D2D cooperation while achieving an outage less than 3% in a low activity scenario. In a high activity scenario, the solution for the optimization problem showed that 9 APs with only 179 LR connections were able to serve 99.95% of the dense network when D2D cooperation is used instead of 66 APs providing 1960 LR connections when no D2D cooperation is considered.

To illustrate the solution for the proposed traffic offloading optimization problem, Figures 5.7, 5.8 and 5.9 show resource allocation of SR and LR connections in $40m \times 40m$ dense area composed of 1960 users (0.7 user activity) for three scenarios: (1) area composed of 5 APs without D2D cooperation, (2) area composed of 5 APs with D2D cooperation, and (3) area composed of 9 APs with D2D cooperation, respectively. As shown in Figure 5.7, 5 APs can only serve 150 LR users which leads to a very high outage percentage of 92.34% with no D2D cooperation due to LR capacity limitation. The outage probability is reduced to 15.81% when D2D cooperation is used (Figure 5.8). 14.54% of users are not served due to capacity limitation (maximum possible capacity is 1650 users) and 1.28% due to coverage limitation. To increase the capacity of the dense network, 9 APs were deployed as shown in Figure 5.9. The outage probability was reduced to 0.05%; only one user was out of coverage.

To study the effect of the user target rate on the performance of the D2D cooperative network, we varied the target rate from 1 to 4 Mbps and evaluated

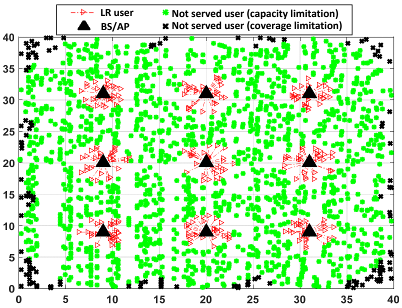


Figure 5.7: Resource allocation for $40m \times 40m$ area composed of 1960 users (0.7 user activity) and 9 APs without D2D cooperation. Outage = 92.34%.

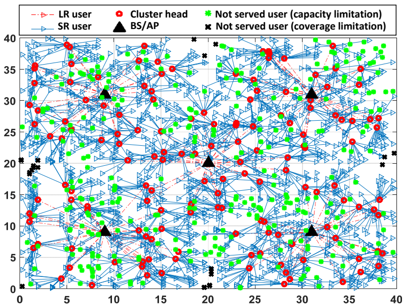


Figure 5.8: Resource allocation for $40m \times 40m$ area composed of 1960 users (0.7 user activity) and 5 APs with D2D cooperation. Outage = 15.81%.

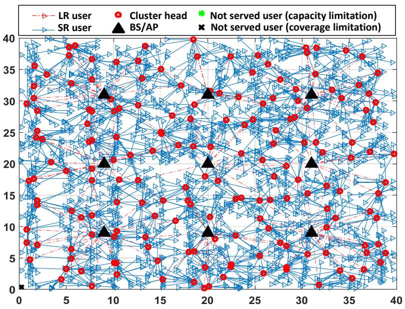


Figure 5.9: Resource allocation for $40m \times 40m$ area composed of 1960 users (0.7 user activity) and 9 APs with D2D cooperation. Outage = 0.05%.

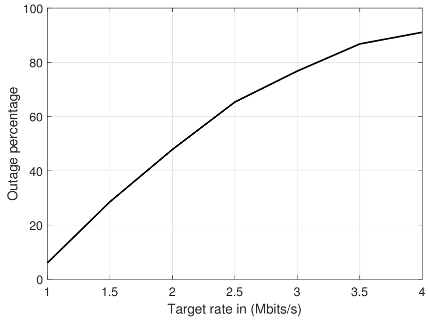


Figure 5.10: Outage percentage variations with respect to user target rate in Mbits/s for $40m \times 40m$ area composed of 280 users and 5 APs.

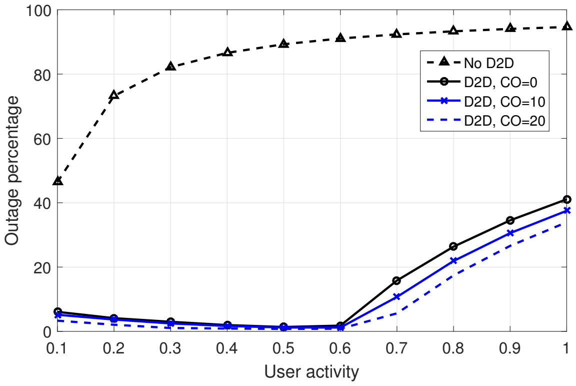


Figure 5.11: Outage percentage variation with user activity probability with/without D2D and 0, 10 and 20 content owners.

the outage probability in a network composed of 5 BSs/APs and 280 users. As shown in Figure 5.10, when the target user rate is low (1Mbps), 17 users were not served which corresponds to an outage percentage of 6.07%. The outage percentage increases with the increase of user target rate to reach 86.78% users not served out of 280.

5.5.4 Performance Results: System Capacity Increase with D2D Cooperation and Content Owners

Figure 5.11 shows the outage probability for different scenarios with and without D2D cooperation while varying the number of users requesting common content simultaneously within $40m \times 40m$ area. The outage probability is very high when no D2D cooperation is considered. The system capacity is limited to 150 possible connections over LR channels when 5 BSs/APs are deployed. This leads to an outage reaching 94.64% when the network density is very high (2800 users).

The outage probability decreases when D2D cooperation is considered. The coverage range and capacity are extended by the cluster heads acting as providers to other MTs. When the network density increases within a specific area, the MTs are closer and tend to use SR channels for data download. Accordingly, the number of clusters formed increases to serve more users. The outage then decreases with the increase of the network density to reach a capacity limited by the number of LR channels K_{L} and the maximum number of users within a cluster K_{C} which is 1650 when 5 APs are deployed. For this reason, the outage probability increases for more than 1680 users. The existence of content owners decreases the outage probability. As presented in Figure 5.11, we considered the existence of 10 and 20 content owners randomly distributed within the $40m \times 40m$

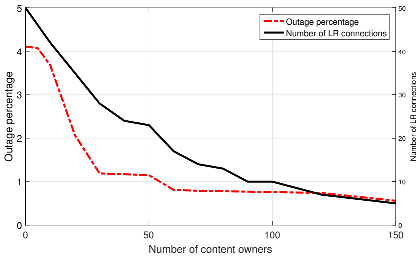


Figure 5.12: The left and right y-axes measure the outage percentage and LR connections variation, respectively, with respect to the number of COs with $K = 560$ users.

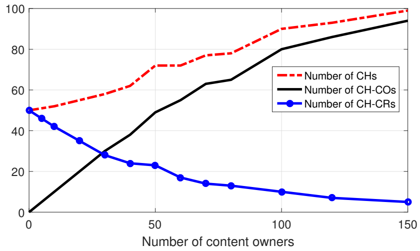


Figure 5.13: The number of clusters CHs, LR users and COs serving as CH-CRs and CH-COs variations with respect to the number of content owners for $K = 560$ users.

area while varying the number of users. The COs served as CHs and transmitted data to other MTs which increased the capacity of the system and reduced the outage probability.

Figure 5.12 shows the outage percentage and LR connections for traffic offloading in D2D cooperative networks where 560 users exist while varying the number of content owners. The coverage range and capacity of the conventional network are extended by the cluster heads acting as providers to other MTs. When no COs are considered, the outage is 4.11%. The outage probability decreases with the increase of the number content owners to be below 1% when 150 COs exist. COs served as cluster heads and transmitted data to other MTs which allow to serve more users.

Figures 5.13 show the number of clusters formed, the number of LR users and COs serving as CHs in D2D cooperative networks where 560 users exist while varying the number of content owners. The number of clusters increases with the increase of the number of COs which expands the network coverage and capacity. The users are encouraged to offload to D2D connections by connecting to a CO rather than connecting to another MT receiving over LR. Accordingly, the outage probability and the number of LR connections decrease.

To illustrate the solution for the proposed traffic offloading optimization problem, Figure 5.14 shows the resource allocation of LR and SR channels in a network composed of 560 users with/without D2D cooperation. The number of LR channels is very high when no D2D cooperation is used. The system capacity is limited to 150 LR connections, which leads to an outage of 410 users (Figure 5.15). The outage probability is reduced to 4.11% when D2D cooperation is used with 50 LR users acting as CH-CRs (Figure 5.16). The number of SR connections increases with the increase of the number of COs, while the number of users not served

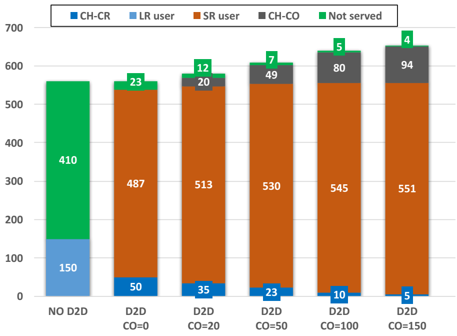


Figure 5.14: LR and SR channel allocation with/without D2D cooperation in a network formed by K=560 users with different number of COs.

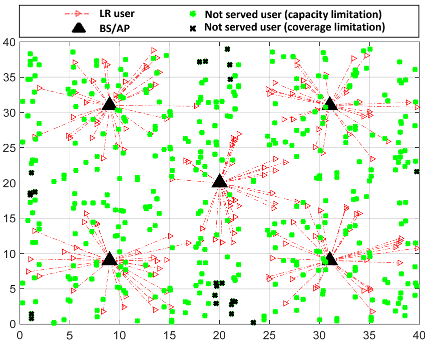


Figure 5.15: Resource allocation without D2D cooperation for $40m \times 40m$ area composed of 560 users and 5 APs. Outage = 73.21%.

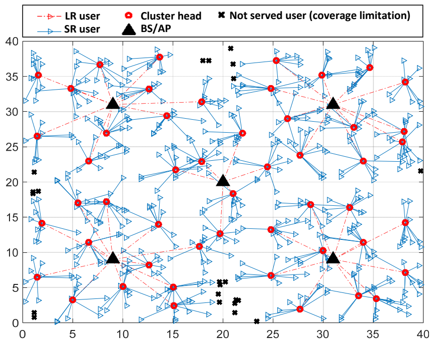


Figure 5.16: Resource allocation with D2D cooperation for $40m \times 40m$ area composed of 560 users and 5 APs. Outage = 4.11%.

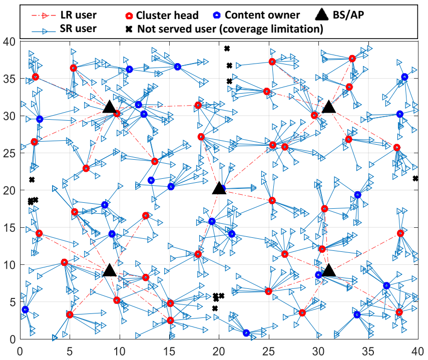


Figure 5.17: Resource allocation with D2D cooperation for $40m \times 40m$ area composed of 560 users, 5 APs and 20 COs. Outage = 2.14%.

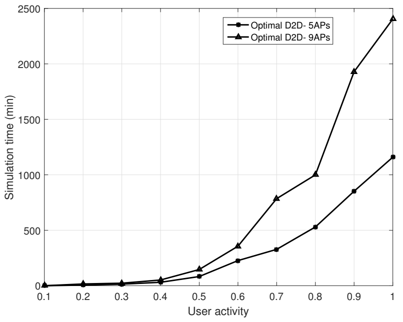


Figure 5.18: Optimization problem solving time (in minutes) variation with user activity probability for $40m\times 40m$ area composed of 5 and 9 APs (2800 users).

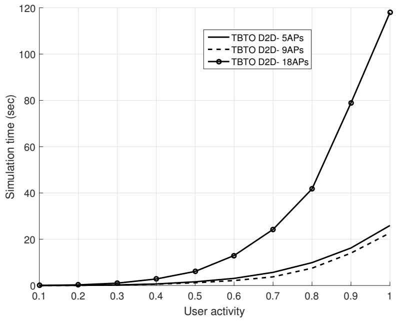


Figure 5.19: TBTO solving time (in seconds) variation with user activity probability for $40m \times 40m$ area composed of 5 and 9 APs (2800 users) and $40m \times 80m$ area composed of 18 APs (5600 users).

decreases. The LR connections are reduced to 35 users while the outage probability was reduced to 2.14% with 20 COs (Figure 5.17). The number of clusters increased from 50 where cluster heads are all LR content recipients CRs, to reach 99 when 150 COs are considered (94 CH-COs and 5 CH-CRs). The number of LR decreases to be only 4 users when 150 COs exist, with 0.71% outage probability.

5.5.5 Performance Results: Proposed Dynamic Tree-Based Traffic Offloading Algorithm

In this section, we compare the optimal traffic offloading solutions with the ones obtained using our proposed tree-based traffic offloading approach. We evaluate the performance in terms of time complexity and user outage percentage. Results demonstrate that the proposed approach TBTO provided real-time and fast solutions with a tradeoff cost in outage probability.

Solving the traffic offloading problem as an optimization problem is computationally expensive. Figure 5.18 shows the solving time in minutes needed by AIMMS to provide solution for optimal traffic offloading in dense heterogeneous networks. The solving time for a low density network (A=0.3) is 14 and 23 min when the network is composed of 840 users, 5 and 9 APs, respectively. As the user activity increases, the solving time increases to reach 5.46 hours and 19.32 hours when the user activity is 0.7 and 1, respectively. Therefore, achieving optimal

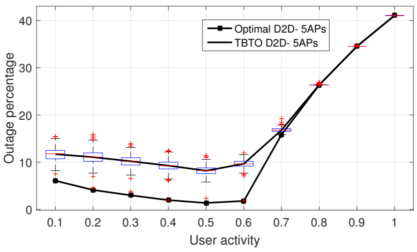


Figure 5.20: Outage percentage variation of the traffic offloading optimal solution and TBTO with user activity probability for $40m \times 40m$ area composed of 5 APs (2800 users).

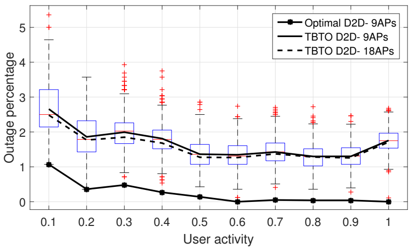


Figure 5.21: Outage percentage variation of the traffic offloading optimal solution and TBTO with user activity probability for $40m \times 40m$ area composed of 9 APs (2800 users), and $40m \times 80m$ area composed of 18 APs (5600 users).

solutions may not be feasible in real-time ultra dense networks.

Our proposed tree-based traffic offloading approach was able to provide near-optimal solutions with very low time complexity. To evaluate the performance of our proposed TBTO approach, we generated solutions for 1000 different order of user arrival. In every simulation, the same distribution of users is used, however, the order of arrival is randomly shuffled. Figure 5.19 shows the time needed to find solutions using the proposed tree-based traffic offloading algorithm. TBTO is able to provide near optimal solutions within few seconds for networks composed of 2800 users and denser networks of 5600 users which was not achievable using optimal traffic offloading. At high user activity (A =1), the simulation time using TBTO was 26 seconds and 23 seconds for a network composed of 2800 users, 5 and 9 APs, respectively. The average simulation time for the networks composed of 9 APs is lower since more users are allowed to be served as LR users, hence, increasing the number of CHs. This increases the possibility of the users to be connected to the tree from the first iteration while satisfying their target rate. Accordingly, the number of delayed users and the number of iterations till all the users are assigned is reduced. Solutions for high user density area of $40m \times 80m$ composed of 5600 users and 18 APs was achievable in maximum of 118 seconds.

Decreasing the simulation time is faced by an acceptable increase in outage probability. Figure 5.20 and 5.21 compares the outage performance of TBTO and optimal traffic offloading for $40m \times 40m$ area composed of 2800 users, 5 and 9 APs, respectively. As presented in Figure 5.20, TBTO provided lower than 8% more outage than optimal traffic offloading in low density networks. In dense

Table 5.4: Performance of TBTO compared to optimal traffic offloading in $40m \times 40m$ area composed on 2800 users and 5 APs

Activity	Optimal traffic offloading		Proposed TBTO approach	
A	**Outage**	**Time**	**Outage**	**Time**
A=**0.1**	8.93%	64ms	6.07%	3.36s
A=**0.2**	4.46%	208ms	4.11%	15.8m
A=**0.3**	3.81%	173ms	2.98%	22.8m
A=**0.4**	1.96%	356ms	1.96%	51.6m
A=**0.5**	2.36%	798ms	1.36%	2.4h
A=**0.6**	1.79%	1.5s	1.79%	5.9h
A=**0.7**	15.82%	3.1s	15.82%	13.1h
A=**0.8**	26.34%	5.6s	26.34%	16.6h
A=**0.9**	34.52%	8.4s	34.52%	32.1h
A=**1**	41.07%	12.5s	41.07%	40.1h

networks, TBTO was able to provide very close outage probability to the optimal solution (A >0.7). Increasing the number of APs to 9 allowed larger number of MTs to be served by the BSs/APs and hence serve as cluster heads and decrease the outage probability. TBTO showed an average of 1.43% increase of outage probability when compared to the traffic offloading optimal solution.

The solution of the proposed approach varies with the arrival order of users since the tree is built sequentially based on the user arrival. Table 5.4 compares for a specific order of user arrival (out of the 1000 shuffles above) the outage and simulation time for TBTO and the optimal traffic offloading algorithm in a network composed of 2800 users and 5 APs. Table 5.4 shows that, for these specific orders, TBTO was able to provide similar outage probability to the optimal solution within few seconds.

5.6 Traffic Offloading Limitations and Challenges

The optimal traffic offloading solutions provide optimal connections and user assignments with minimum long range users. However, due to the complexity of the problem, optimal solutions may not be achievable for ultra dense D2D cooperative networks where a large number of users are requesting data simultaneously. Towards this goal, we proposed a dynamic tree-based traffic offloading approach which assigns the users' connections sequentially based on a tree having BSs/APs and mobile terminals as nodes. The order of arrival of users affect the performance of the TBTO proposed approach. The proposed approach provided real-time and fast solutions even in very high density user activity. The cost of reducing the time complexity is faced by an acceptable increase in user outage probability.

The proposed TBTO approach presented solution for traffic offloading while considering orthogonal channel allocation. However, in real network scenarios, limited number of orthogonal channels are available which leads to high channel reuse and degradation of service. The sub-optimal traffic offloading approach needs then to consider simultaneously the allocation of non-orthogonal channels to reflect real network scenarios.

Chapter 6

Non-Orthogonal Channel Allocation for Minimum Interference in Dense D2D Cooperative Networks

In Chapter 5, we addressed resource management in dense D2D cooperative heterogeneous networks to determine the optimal traffic offloading strategies while all the channels allocated to cluster heads are assumed to be orthogonal.

In this chapter, we address non-orthogonal channel allocation using the solution obtained from the traffic offloading problem. We aim at allocating the available orthogonal SR channels to the mobile terminals serving as cluster heads while reducing the interference caused by channel reuse. We first formulate the channel allocation as an optimization problem aiming at minimizing the reuse of the channels and maximizing the distance between non-orthogonal channels to reduce the interference. We then solve the optimization problem using AIMMS software and CPLEX as a solver. Moreover, we evaluate the effect of optimal non-orthogonal channel allocation to cluster heads. Allocating non-orthogonal channels affects the service transmission rate and may lead to service degradation below the target rate. Therefore, we propose an enhanced user allocation scheme to serve the affected users by assigning them either to existing cluster heads or BSs/APs.

Due to the complexity of the channel allocation optimization problem, optimal solutions may not be achievable in real-time ultra dense D2D cooperative networks. For this reason, we propose a dynamic tree-based resource management (TBRM) approach that includes hierarchical traffic offloading and channel allocation simultaneously. We used similar approach presented in Chapter 5 where the users' connections are assigned consecutively based on a tree having BSs/APs and mobile terminals as nodes. However, we consider simultaneously allocating available channels to cluster heads aiming at maximizing the distance

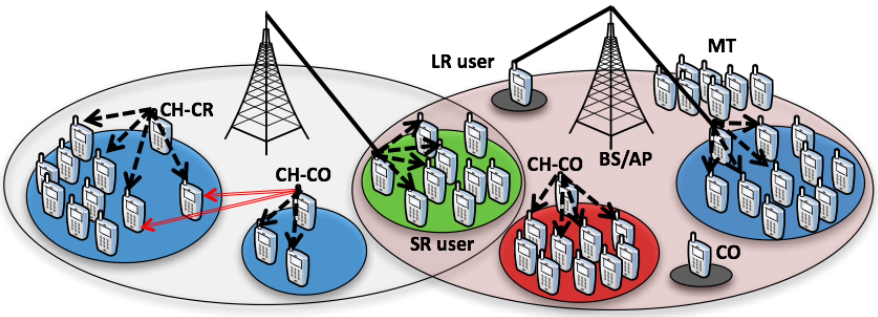


Figure 6.1: D2D cooperative network formed by one BS/AP and five clusters. Three cluster heads use the same channel which causes interference to SR users.

between mobile terminals using co-channels to reduce interference. To evaluate our solutions, we focus on a stadium topology to demonstrate the significant gains of optimized non-orthogonal channel allocation in ultra dense wireless networks. We show that the proposed approach is able to provide near-optimal solutions with very low time complexity.

This chapter is organized as follows. The system model is presented in Section 6.1. The optimal channel allocation for minimum interference in dense D2D cooperative networks is presented in Section 6.2. Enhanced allocation approach to assign connections to the users affected by non-orthogonal channel allocation is proposed in Section 6.3. Dynamic tree-based resource management approach is presented Section 6.4. Performance results are presented in Section 6.5. Resource management limitations and challenges are presented in Section 6.6.

6.1 System Model

In this chapter, we use similar system model presented in Chapter 5. We consider a network formed by M BSs/APs and a large number of mobile terminals K. A MT i requests a common content data such as live and on-demand video streaming with a specific transmission target rate $R_{\mathrm{T},i}$. MT i may receive its data from a BS/AP over LR wireless technology or from another MT cluster head using SR wireless technology. A cluster head can be either:(1) a content owner CH-CO which is a mobile terminal who has already the data content cached, or (2) a content recipient CH-CR who is a LR user receiving data from BS/AP over LR channel. SR channels are allocated to cluster heads to serve the users within a cluster.

The number of orthogonal channels become scarce with the increase of number of users in dense device-to-device cooperative networks. For instance, 2.4GHz

WLAN IEEE 802.11n provides three non-overlapping channels 1, 6 and 11. 5GHz IEEE 802.11n provides three Unlicensed National Information Infrastructure (UNII) bands: (1) UNII-1 containing channels four orthogonal channels (36, 40, 44, and 48), (2) UNII-3 containing 4 orthogonal channels (149, 153, 157 and 161), and (3) UNII-2 containing 15 channels (52, 56, 60, 64, 100, 104, 108, 112, 116, 120, 124, 128, 132, 136, and 140) shared with radar systems [119].

Allocating same channel to multiple cluster heads causes interference to the SR users receiving from these cluster head mobile terminals. As presented in Figure 6.1, the network is formed by BSs/APs, content owners and clusters served by a cluster head mobile terminal. A MT is subjected to interference from cluster heads using non-orthogonal channels while receiving over SR. In our case model, we consider a network composed by a limited number of base stations or access points. We then assume that all the LR channels allocated are orthogonal. Therefore, the LR mobile terminals will not be subjected to LR interference. However, the SR users will be subjected to interference caused by cluster heads using non-orthogonal channels.

6.2 Optimal Channel Allocation for Minimum Interference Formulation

In this section, we formulate the SR channel allocation problem as an optimization problem aiming at serving the maximum number of users while reducing the effect of interference caused by channels reuse. The goal is to allocate the available non-overlapping channels to the cluster heads assigned by the optimal traffic offloading solution. Reducing interference can be achieved by assigning orthogonal channels with minimum reuse to distant cluster heads. Therefore, the channel allocation problem aims at minimizing the channels reuse and maximizing the distance between non-orthogonal channels.

Accordingly, the decision variable $\mathbf{Q}$ is considered to indicate the channel allocation (Table 6.1). The binary decision variable Q_{pi} indicates whether the channel p is allocated to mobile terminal cluster head i to transmit the data to other mobile terminals. The decision variable $\mathbf{Q}$ will then be a matrix of dimension $N_{CH} \times N_{SRo}$ where N_{CH} is the number of cluster heads considered in the network and N_{SRo} is the maximum number of orthogonal SR channels.

- Q_{pi}: binary variable that indicates whether channel p is allocated to mobile terminal i.

$$Q_{pi} = \begin{cases} 1 & \text{if channel } p \text{ is allocated to cluster head } i \\ 0 & \text{otherwise} \end{cases} \tag{6.1}$$

The channel allocation problem can be formulated as follows:

Table 6.1: Main parameters and variables

Parameters	
K	the set of requesting MTs, where a MT is referred to as MT i, $i = 1, ..., K$
M	the set of BSs/APs, where a BS/AP is referred to as BS/AP m, $m = 1, ..., M$
N_{SRo}	number of orthogonal non-overlapping SR channels
d_{ij}	distance between transmitter (BS/AP or MT) i and MT j
$R_{\mathrm{S},ij}$	transmission rate on SR from the MT i to MT j
$R_{\mathrm{T},i}$	target transmission rate to MT i to meet video application requirements
N_{LR}	maximum number of LR channels in the network
N_{SR}	maximum number of clusters in the network
N_{CH}	number of cluster heads in the network
$K_{\mathrm{L},m}$	maximum number of MTs served by a BS/AP m
$K_{\mathrm{C},i}$	maximum number of MTs served over SR by a cluster head MT i in a cluster
Variables	
Q_{pi}	a binary variable that indicates whether channel p is allocated to cluster head i

$$\underset{\mathbf{Q}}{\text{argmax}} \quad \sum_{i=1}^{N_{CH}} \sum_{j=1}^{N_{CH}} \sum_{p=1}^{N_{SRo}} d_{ij} \cdot Q_{pi} \cdot Q_{pj} \tag{6.2}$$

subject to

$$\sum_{p=1}^{N_{SR}} Q_{pi} = 1, \forall i \tag{6.3}$$

$$\left\lfloor \frac{N_{CH}}{N_{\mathrm{SRo}}} \right\rfloor \leq \sum_{i=1}^{N_{CH}} Q_{pi} \leq \left\lceil \frac{N_{CH}}{N_{\mathrm{SRo}}} \right\rceil, \forall p \tag{6.4}$$

$$Q_{qi} \in \{0, 1\} \tag{6.5}$$

where

- Constraint (6.3) ensures that every cluster head mobile terminal is assigned one SR channel.
- Constraint (6.4) ensures that all the channels are used with minimum reuse factor. If the number of orthogonal channels N_{SRo} is greater than the number of cluster heads N_{CH}, every channel p can be then assigned maximum once $\left(0 \leq \sum_{i=1}^{N_{CH}} Q_{pi} \leq 1\right)$. If the number of orthogonal channels is less than the number of CHs, constraint (6.4) ensures that a channel p is allocated with minimum reuse, i.e. maximum of $\left\lceil \frac{N_{CH}}{N_{\mathrm{SRo}}} \right\rceil$ and minimum of $\left\lfloor \frac{N_{CH}}{N_{\mathrm{SRo}}} \right\rfloor$.
- The last constraint sets the decision variable $\mathbf{Q}$ to be binary.

6.2.1 Solution Methodology

The problem is a binary linear programming problem. The number of binary variables is $N_{CH} \times N_{SRo}$ composed of the decision variable $\mathbf{Q}$ indicating the channel allocation to cluster heads. The problem is NP-complete. Starting with the first constraint, the problem is divided into multiple directions or directed graphs based on a channel allocation to a cluster head mobile terminal. Each graph is composed of different subgraphs based on the problem constraints. The problem is similar to Minimum Dominating Set problem in Directed Graphs which has been shown as NP-Hard [115][116]. It can be shown that solution for the optimal channel allocation problem can be verified in polynomial time, thus the problem is NP-complete.

6.3 Proposed Enhanced User Allocation Approach in Dense D2D Cooperative Networks

Allocating non-orthogonal channels to cluster heads affects the system performance. The transmission service rate decreases due to the interference caused by channel reuse. Even if the channels are optimally allocated to cluster head mobile terminals, SR users are subjected to interference which may degrade the service rate below the target service rate. Accordingly, we evaluate the effect of allocating the available orthogonal channels to the cluster heads on the SR users. We then present possible solutions for allocating connections to the users affected either to other mobile terminals serving as cluster heads or BSs/APs.

As presented in Algorithm 6.1, we check possible assignment solutions for every SR user i whose target rate is no longer met due to channel reuse interference. We start by checking the transmission rates provided by every cluster head in the network taking into consideration the interference caused by cluster heads using same channels. The MT i will be then assigned to the cluster head n satisfying MT i target rate and serving lower number of users (below $K_{\text{C},n}$). In case, there is no cluster head satisfying system constraints, the LR transmission rates are examined. MT i can be assigned to the BS/AP m satisfying MT i target rate and serving lower number of users (below $K_{\text{L},m}$). If the system constraints are not satisfied, the MT i is considered not served.

6.4 Dynamic Tree-Based Resource Management in Dense D2D Cooperative Networks

Optimal solutions for traffic offloading and channel allocation may not be achievable in real-time ultra dense D2D cooperative networks. In additional, allocating channels based on the solution provided by the traffic offloading assuming the

Algorithm 6.1: The proposed enhanced user allocation approach

```
Input:
- $K$          number of users,
- $M$          number of BSs/APs,
- $d_{ij}$     distance between transmitter (BS/AP or MT) $i$ and MT $j$,
- $N_{\mathrm{LR}}$     maximum number of LR channels in the network,
- $N_{\mathrm{SR}}$     maximum number of SR channels in the network,
- $K_{\mathrm{L},m}$    maximum number of MTs served by a BS/AP $m$,
- $K_{\mathrm{C},i}$    maximum number of MTs served by a cluster head MT $i$,
- $R_{\mathrm{T},i}$    target transmission rate to MT $i$ to meet video application requirements,
- $\mathbf{z}$          users served determined by the optimal traffic offloading solution,
- $\mathbf{y}$          LR connections determined by the optimal traffic offloading solution,
- $\mathbf{v}$          LR connections determined by the optimal traffic offloading solution,
- $\mathbf{u}$          cluster heads determined by the optimal traffic offloading solution,
- $\mathbf{Q}$          channel allocation by the optimal channel allocation solution

Output:
- $\mathcal{Z}_i$       a binary variable that indicates whether MT $i$ is receiving data,
- $\mathcal{Y}_{mi}$    a binary variable that indicates whether MT $i$ is receiving data over LR from BS/AP $m$,
- $\mathcal{V}_{ij}$    a binary variable that indicates whether MT $j$ is receiving data over SR from MT $i$

Initialize $\mathcal{Z}, \mathcal{Y}$ and $\mathcal{V}$ to be equal to $\mathbf{z}$, $\mathbf{y}$ and $\mathbf{v}$, respectively
Compute SR user $i$ transmission rate $R_i$ achieved after optimal channel allocation considering the interference caused by cluster heads using non-orthogonal channels $R_i$
Assign a new connection to user $i$ if transmission rate $R_i$ is lower than target rate as follows:
    1. Estimate the transmission rate $R_{\mathrm{S},ni}$ provided by a cluster head $n$ considering the interference caused by all the cluster heads using the same channel of cluster head $n$.
    2. Consider every cluster head $n$ providing transmission rates higher than the target rate of MT $i$ and serving less than $K_{\mathrm{C}}, n$ users, as candidate cluster heads $CCH$.
       $CCH = \{n | n \in \mathcal{C}, R_{\mathrm{S},ni} \geq R_{\mathrm{T},i}$ & $\sum_{j=0}^{K} v_{nj} < K_{\mathrm{C},n}\}$
    3. Check if SR and LR transmission rates satisfy MT $i$ requirements
        • if $CCH = \emptyset$ then
        •     Consider MT $i$ no longer served as SR user and set $V_{xi} = 0, \forall x$
        •     Check if MT $i$ can be served over LR links
                  – Consider every BS $m$ providing transmission rates higher than the target rate of MT $i$ and serving less than $K_{\mathrm{L}}, m$ users, as candidate base station $CBS$.
                    $CBS = \{m | m \in \mathcal{M}, R_{\mathrm{L},mi} \geq R_{\mathrm{T},i}$ & $\sum_{j=0}^{K} y_{mj} < K_{\mathrm{L},m}\}$
```

```
2:    3.    •    – if $CBS = \emptyset$ then
                 –     Consider MT $i$ a non-assigned users and set $Z_i = 0$
                 – else
                 –     Assign MT $i$ to BS/AP $m$ with lower number of users
                       served and set $Y_{mi} = 1$
                 – end if
            • else
            •     Assign MT $i$ to cluster head $n$ serving lower number of users
                  and set $V_{ni} = 1$
            • end if
3: Repeat process (2) for all the SR users.
```

channels are orthogonal may lead to high user outage and may not be feasible in real-time networks. Moreover, the optimization problem is holistic and considers all the existing users when providing optimal solutions, which may not be feasible in real-time networks. Accordingly, providing sub-optimal solutions for traffic offloading with channel allocation considerations are needed to provide a balance between time complexity, user outage, number of LR channels. For these reasons, we propose a dynamic tree-based resource management (TBRM) approach that includes hierarchical traffic offloading and channel allocation simultaneously.

The TBRM approach is based on the TBTO approach presented in Chapter 5 where the users' connections are assigned consecutively based on a tree having BSs/APs and mobile terminals as nodes. However, TBRM considers simultaneously allocating the available channels to cluster heads consecutively while maximizing the distance between co-channels. The interference caused by co-channel reuse is then considered while assigning SR links to the users.

The proposed dynamic tree-based resource management approach is based on a four-level tree as follows: (1) the network as root node, (2) the BSs/APs as first level parent nodes, (3) the cluster heads and LR users as second level nodes receiving data from BSs/APs, and (4) SR users receiving data from cluster heads as fourth level terminal nodes. In general, a mobile terminal can be connected to a BS/AP or another MT serving as cluster head to download common content data. Accordingly, we consider BSs/APs or CHs as parent nodes. When connected to a BS/AP, a MT j is considered a LR user and is added to the tree. When connected to another mobile terminal i, MT j is added to the tree in the fourth level as a SR user. In addition, the mobile terminal i becomes a cluster head and is assigned a channel. Accordingly, every BS/AP forms its own sub-tree having LR users and cluster heads as their child nodes. Similarly, the CHs form their own sub-trees, called clusters in our model, having SR users as child nodes. We define $\mathcal{N}$ as the set of existing nodes in the tree composed of BSs/APs and assigned users.

To assign users' connections, the algorithm builds a tree considering users sequentially, one user at a time. It starts from the root node which is in our case the network. The root has by default all the BSs/APs as child nodes. To assign the connection for MT j, the nodes $\mathcal{N}$ of the current tree are considered, and the mobile terminal j is added as a child node to the tree based on Algorithm 6.2. In our work, we aim at maximizing network capacity by minimizing the number of LR channels used, which allows to serve more users. In order to increase the capacity of the network, the mobile terminals are encouraged to use D2D connectivity. Accordingly, we gave twice the weights of LR connectivity to D2D SR connectivity.

As presented in Algorithm 6.2, the connection for a mobile terminal j starts by estimating the transmission rates provided by existing nodes $\mathcal{N}$: BSs/APs (level 2 nodes) and LR users and cluster heads (level 3 nodes). The SR transmission rate $R_{\mathrm{S},nj}$ provided by another mobile terminal n over SR link is estimated taking into account the interference caused by the set of cluster heads $\mathscr{C}_n$ using same channel as MT n. The mobile terminal j can be served by a cluster head which has already a channel p assigned, or a LR user n which needs a SR channel allocation. To perform channel allocation for LR MT n, we aim at maximizing the distance between mobile terminals using co-channels to reduce interference. Accordingly, for every channel p, we compute the total distance between the MT n and cluster heads using the same channel p. The channel providing the maximum distance is then selected to be allocated to LR user n.

The nodes providing transmission rates higher than the target rate of MT j and serving less than the maximum number of allowed users, are considered as candidates nodes. The weight of SR rate $R_{\mathrm{S},nj}$ between the MT n and MT j is assumed to be twice the weight of the LR rate $R_{\mathrm{L},mj}$ between the BS/AP m and MT j to encourage D2D traffic offloading. The candidate node providing higher system throughput is then selected. In case two nodes provide the same throughput, the mobile terminal j will be connected to the node serving less MTs. The assignment also makes sure the system constraints are satisfied such as the maximum number of users in a cluster, the maximum number of allowed LR users and SR users. If the target rate of a MT j is not satisfied by the nodes in the current tree, the mobile terminal assignment are delayed and reconsidered later. This process continues until all users are considered and no more users can be added to the tree.

6.5 Performance Results and Analysis

In this section, we present our simulation setup and performance results and analysis. We first present performance results for optimal channel allocation while varying the density of the network and the number of available non-orthogonal channels. We then present possible user allocation as enhancement to the per-

Algorithm 6.2: The proposed dynamic tree-based resource management (TBRM) approach

Input:

- K number of users,
- M number of BSs/APs,
- d_{ij} distance between transmitter (BS/AP or MT) i and MT j,
- N_{LR} maximum number of LR channels in the network,
- N_{SR} maximum number of SR channels in the network,
- N_{SRo} maximum number of SR orthogonal non-overlapping channels,
- $K_{\mathrm{L},m}$ maximum number of MTs served by a BS/AP m,
- $K_{\mathrm{C},i}$ maximum number of MTs served by a cluster head MT i,

Output:

- z_i a binary variable that indicates whether MT i is receiving data
- y_{mi} a binary variable that indicates whether MT i is receiving data over LR from BS/AP m
- v_{ij} a binary variable that indicates whether MT j is receiving data over SR from MT i
- u_i a binary variable that indicates whether MT i is a cluster head

```
1: Consider the network as root node
2: Assign BSs/APs as default child nodes to the network root node. The BSs/APs
   will be then considered first level nodes in the tree. At this stage, $\mathcal{N} = \{m | m \in M\}$
3: Assign a connection and allocate a parent node for a MT $j$ as follows:
       1. Estimate the transmission rate $R_{\mathrm{L},mj}$ over LR from a BS/AP $m$.
       2. Estimate the transmission rate $R_{\mathrm{S},nj}$ over SR from a cluster head $n$ existing
          in the tree $\mathcal{N}$, which is assigned channel $p$.
       3. Estimate the transmission rate $R_{\mathrm{S},nj}$ over SR from a LR user $n$ existing
          in the tree $\mathcal{N}$, which is not assigned any channel yet. Accordingly, channel
          allocation should be performed first to allocate a channel $p$ to MT $n$ while
          aiming at maximizing the distance between user using the same channel $p$
          to reduce interference as follows:
              • Compute the distance between LR user $n$ and the set of cluster
                heads using the same channel for every channel $p$
              • Select the channel providing the higher sum of the distances as
                potential channel to be allocated to LR user $n$ if MT $j$ is assigned
                to MT $n$ at the end
       4. Consider the nodes providing transmission rates higher than the target
          rate of MT $j$ as candidate nodes CN.
          $CN = \{n | n \in \mathcal{N}, R_{\mathrm{X},nj} \geq R_{\mathrm{T},j}\}$
```

```
3:    5. Check if transmission rates satisfy MT $j$ target rate
          • if $CN = \emptyset$ then
          •     Delay MT $j$ assignment until all users are considered
          •     Add MT $j$ to the non-assigned users
          • else
          •     Consider the weight of SR rate $R_{\mathrm{S},nj}$ between the existing CH
                and LR users nodes $n$ and MT $j$ twice the weight of the LR rate
                $R_{\mathrm{L},nj}$ between the BSs/APs nodes $n$ and MT $j$ to encourage traffic
                offloading to D2D connectivity
          • Check system constraints:
                – if the candidate node $n$ is a CH MT $i$ & the number of its
                  child nodes $= K_{\mathrm{C},i}$ then Eliminate node $n$ end if
                – if the candidate node $n$ is a BS/AP $m$ & the number of its
                  child nodes $= K_{\mathrm{L},m}$ then Eliminate node $n$ end if
                – if the candidate node $n$ is a BS/AP $m$ & the number of LR
                  users $= N_{\mathrm{LR}}$ then Eliminate node $n$ end if
                – if the candidate node $n$ is a MT $i$ & the number of CHs $= N_{\mathrm{SR}}$
                  then Eliminate node $n$ end if
          • Select the candidate node $n$ providing higher system throughput
                – if multiple nodes provide the same system throughput then
                  Select node $n$ serving less child nodes end if
          • Add the MT $j$ as a child to node $n$
          • Update the tree, $\mathcal{N}$ and decision variables:
                – if MT $j$ is added to the tree then MT $j$ is served & $z_j$ is set
                  to 1 end if
                – if node $n$ is a BS/AP $m$ then MT $j$ is a LR user & $y_{mj}$ is
                  set to 1 end if
                – if node $n$ is a MT $i$ then
                –     MT $j$ is a SR user & $v_{ij}$ is set to 1
                –     if node $n$ is a LR user $i$ then
                –         MT $i$ is a CH & $u_i$ is set to 1
                –         Selected channel $p$ is allocated to LR user $i$
                –     end if
                – end if
          • end if
4: Repeat process (3) for all the non-assigned users until no more users can be
   added as nodes to the tree.
```

Table 6.2: Network parameters and assumptions

Parameters	Values
Section area	40m × 40m
Seats capacity C	2800 seats/section
M	5
A	0.1 - 1
K	$C \times A$
$R_{\mathrm{T},i}$	1 Mbps
$K_{\mathrm{L},m}$ (K_{L})	30 connections/AP m
$K_{\mathrm{C},i}$ (K_{C})	10 connections/CH i
N_{LR}	$\sum_{m=1}^{M} K_{\mathrm{L},m}$
N_{SR}	$N_{\mathrm{LR}} \cdot K_{\mathrm{C}}$
P_{tLR}	10 Watts
P_{tSR}	0.5 Watts
W	0.5 MHz
P_{e}	10^{-3}
σ^2	10^{-3} Watts
κ	-31.54 dB
α	3.71
d_0	10 m

formance of optimal allocation of non-orthogonal channels using our proposed enhanced user allocation approach. We also present performance results for our proposed dynamic tree-based resource management considering traffic offloading and channel allocation simultaneously.

6.5.1 Simulation Setup

We use similar simulation setup presented in Chapter 5. We consider a stadium with a capacity of 100,000 seats. Due to the high complexity of the problem and the large number of users, we divided the area into small sections of $40m \times 40m$. The main system parameters are summarized in Table 6.2.

6.5.2 Performance Results: Optimal Channel Allocation

Figure 6.2 shows the outage probability for optimal channel allocation in a network composed of 5 BSs/APs and 280 users, while varying the number of available orthogonal non-overlapping channels. The solution of the optimal channel allocation is based on the solution provided by the traffic offloading optimization problem, which serves in this case 263 users, 37 LR users serving as cluster heads, 226 SR users and an outage of 17 users (6.07%).

The outage probability decreases with the increase of the number of the available channels since the channels are less reused. When three orthogonal channels are used, channels 1, 2 and 3 are reused 12, 12 and 13, respectively, to be assigned to 37 cluster heads. 88 SR users are affected by the interference caused by

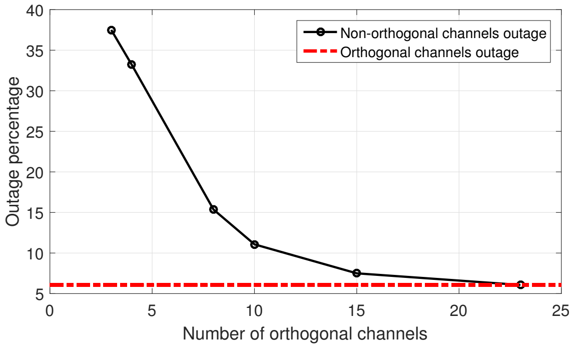


Figure 6.2: Outage percentage variation in a network composed of 280 users (37 cluster heads) with respect to available orthogonal channels.

channel reuse as presented in Figure 6.3, which leads to an additional outage of 31.42%. The total outage will be 37.49%. When the number of available channels used increases to 8, the outage due to co-channel allocation is reduced to 9.28% (Figure 6.4). The outage decreases with the increase of number of available non-overlapping channels to be 0% when 23 channels are used. In this case, the channels are reused maximum twice and the mobile terminals using same channels are distant that the interference caused is negligible.

6.5.3 Performance Results: Proposed Enhanced User Allocation Approach

Figures 6.5(a) and 6.5(b) compare the outage percentage and number of LR channels, respectively, while varying the user density for different scenarios: (1) orthogonal channel allocation which is the solution of the optimal traffic offloading, (2) three orthogonal channel allocation, and (3) the proposed enhanced user allocation approach, and (4) proposed TBRM approach. The outage percentage considering three orthogonal channel allocation increases from 37.5% at low user density of A=0.1 (280 users) to reach 88.46% at high user density of A=1 (2800 users). The proposed enhanced user allocation approach was able to provide lower outage probability with a tradeoff cost in LR channels. The outage was reduced from 37.5% to 12.14% with an additional use of 58 LR channels for low density networks (280 users). It was reduced from 88.46% to 57.39% with a limited number of LR channels of 150. To illustrate a solution, Figure 6.6 shows the additional LR and SR connections assigned by the enhanced proposed approach to serve higher number of users in a network formed by 280 users. Due to the low number of available orthogonal non-overlapping channels, the target service rate was not satisfied for a large number of users which limits the D2D offloading

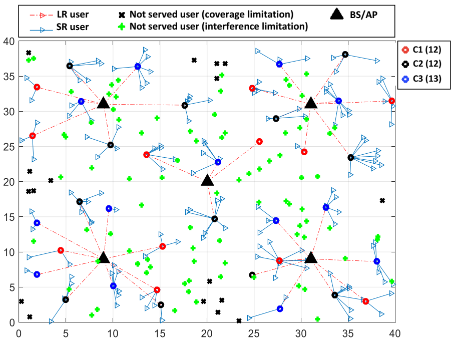

LR user
SR user
Not served user (coverage limitation)
Not served user (interference limitation)
BS/AP
C1 (12)
C2 (12)
C3 (13)
0
5
10
15
20
25
30
35
40

Figure 6.3: Channel allocation of 3 orthogonal channels for $40m \times 40m$ area composed of 280 users and 5 APs. Orthogonal channel allocation outage= 6.07%. Additional outage= 31.42%.

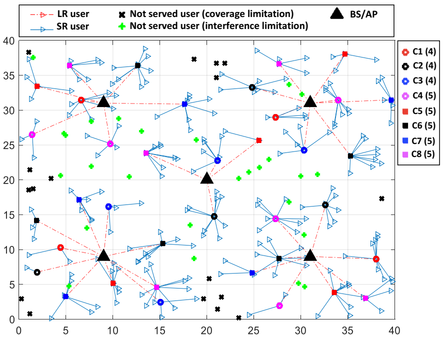
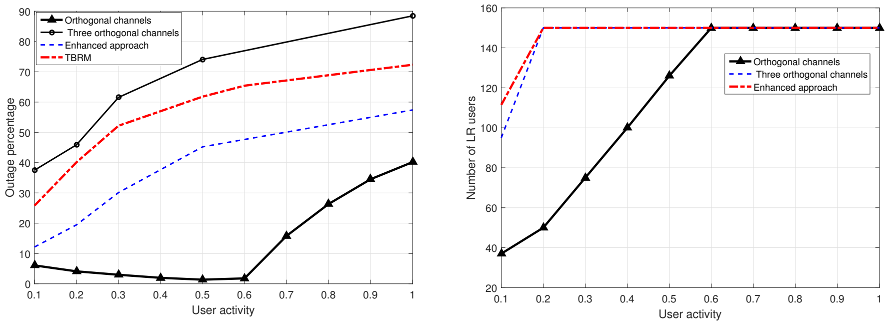


Figure 6.4: Channel allocation of 8 orthogonal channels for $40m \times 40m$ area composed of 280 users and 5 APs. Orthogonal channel allocation outage= 6.07%. Additional outage= 9.28%.



(a) Outage percentage

(b) Number of LR users

Figure 6.5: Performance variation with different user activity for the following scenarios: (1) optimal traffic offloading with orthogonal channel allocation, (2) three orthogonal channel allocation, (3) enhanced proposed approach, and (4) proposed TBRM approach.

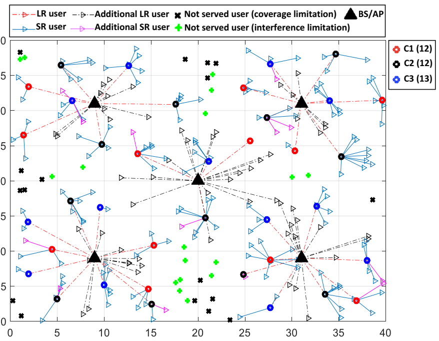


Figure 6.6: Enhanced user allocation approach using 3 orthogonal channels for $40m \times 40m$ area composed of 280 users and 5 APs. Outage= 12.14%.

and force the usage of all the LR channels.

6.5.4 Performance Results: Proposed Dynamic Tree-Based Resource Management Approach

In this section, we generated results for 100 runs of the proposed dynamic tree-based resource management approach while varying the users' arrival order. Figure 6.4(a) compares the outage probability of the proposed TBRM approach with optimal channel allocation with/without orthogonal channels considerations and enhanced proposed solution. The outage caused by optimal channel allocation based on the solution of the traffic offloading problem showed high outage probability. The outage was reduced by the enhanced proposed approach with a tradeoff cost in LR channels as shown in Figure 6.4(b). The proposed TBRM approach was able to allocate non-orthogonal channels simultaneously while performing traffic offloading with notably low time complexity with a tradeoff cost in terms of outage probability and LR channels. The proposed tree-based resource management provides higher outage probability than the enhanced proposed approach, however, less than the optimal channel allocation based on traffic offloading optimal solution. The number of LR users was higher at low network density and reaches the maximum LR users capacity when the number of users is above 560. This shows that our proposed TBRM approach was able to provide dynamic sub-optimal solutions with very low time complexity.

6.6 Resource Management Limitations and Challenges

In this chapter, we presented optimal non-orthogonal channel allocations. We evaluated the effect of the optimal channel allocations on the SR users. The results showed high outage probability which can be improved by assigning the users affected to other cluster heads or BSs/APs. In addition, we propose a dynamic tree-based resource management approach to simultaneously consider traffic offloading with non-orthogonal channel allocation and provide sub-optimal solutions with low time complexity in real-time dense networks.

The solution of the traffic offloading and channel allocation is first limited by the maximum number of users served by BS/AP or by a cluster head. Increasing the BS/AP transmit power may be a solution for enhancing system coverage, however, it causes higher interference to users and service degradation. Accordingly, enhanced power allocation schemes may be proposed as a solution for system capacity and coverage limitations providing a tradeoff between coverage, capacity and interference. In addition, we assume in our work that all the MTs are willing to act as cluster heads, however, usually MTs are selfish and need incentives to participate in the cooperative content distribution.

Chapter 7

Conclusion

This chapter summarizes the main contributions of this dissertation in Section7.1. It also highlights connections to existing wireless standards that discuss device-to-device cooperation, and the inter-operation between different radio access technologies in Section 7.2. Finally, selected open research directions are presented in Section 7.3.

7.1 Contributions

In this dissertation, we addressed user-centric strategies for resource management in heterogeneous networks with quality of service considerations. Our research work is divided between two directions: the first direction for user-centric cellular/WiFi resource management strategies for a single user, and the second direction for multi-user resource management including traffic offloading and channel allocation in dense cache-enabled D2D cooperative heterogeneous networks.

Towards the first direction, we present static cellular/WiFi and traffic splitting for data download. The mobile terminal is assumed to be equipped with two wireless interfaces (cellular 3G/4G network and WiFi). The main challenge is to determine the amount of data to be downloaded over each interface to achieve performance gains. We then extended our work to consider dynamic cellular/WiFi traffic splitting for video streaming with quality of experience considerations. In this case, real-time decisions are needed every time slot duration on the amount of data to be downloaded over cellular and WiFi interfaces maximizing quality of experience while reducing energy consumption and delay.

Towards the second direction, we solve multi-user traffic offloading in dense cache-enabled device-to-device cooperative heterogeneous networks, where a very high number of users request simultaneously a huge amount of data using a stadium case model. The challenging point is to determine which MTs should receive the content on the long range interface and to which MTs they should transmit the received content on the short range interfaces. Moreover, the MTs that receive on SR can, in turn, act as relays to transmit their received content to other MTs on the SR interfaces. We focus on finding optimized strategies for common data download to serve the maximum possible number of users while maintaining a target user quality of service. Due to high complexity of the problem, sub-optimal approaches are proposed for real-time and fast resource management including traffic offloading and non-orthogonal channel allocation.

The main contributions of this dissertation can be summarized as follows:

1. Proposing a learning-based user-centric approach for performing static network selection based on real-network implementations. In contrast to literature, the model considers the features that affect the selection decision known by the user: availability of the networks, signal strength reflecting the channel quality, data size, battery life, speed of the user, location, and type of application. We present an approach for building training data as a basis for machine learning of network selection and then develop decision-tree classification model for network selection that provides the user either the highest quality of service, lowest energy consumption or highest energy efficiency based on pre-defined rules.

2. Developing the static user-centric traffic splitting as an optimization problem, where a mobile device can simultaneously utilize both wireless interfaces. The aim is to guarantee a balance between energy consumption and throughput based on the user's needs in terms of application service requirements and mobile device battery life. Moreover, experimental measurements are used to determine values for key parameters in order to evaluate the proposed traffic splitting approach under realistic network conditions.

3. Proposing dynamic user-centric traffic splitting with delay-power-QoE balance for real-time video streaming applications. We developed an optimized multi-objective traffic splitting solution that minimizes delay, stabilizes network queue while reducing energy consumption and achieving high quality end-user experience. In contrast to the literature, our approach does not focus only on throughput and energy consumption, but also considers user quality of experience based on ITU-T P.1201 standard to better capture the video quality as perceived by the end user. Lyapunov drift plus penalty optimization approach is used to provide real-time traffic splitting decisions as a function of the dynamic variation of various system parameters while achieving a balance between high quality end-user experience, low energy consumption and delay, and allowing the use of multiple interfaces simultaneously to split the traffic at a specific time slot into different links. The proposed approach was evaluated under realistic operational conditions using our own test bed, where the proposed approach is implemented as an Android application that functions in the background at the user side without any intervention from the network or the server, and without performing

any changes to the cellular/WiFi standards.

4. Formulating the multi-user resource management, considering simultaneously traffic offloading and non-orthogonal channel allocation, as an optimization problem aiming at maximizing the user capacity in dense cache-enabled device-to-device heterogeneous networks. The complexity of the problem is studied and reduced to address first optimal traffic offloading problem assuming all the channels are orthogonal, and then optimal non-orthogonal channel allocation. The optimal traffic offloading in dense cache-enabled D2D cooperative network is formulated as an optimization problem to find the optimal LR and SR channels allocation constrained by the number of APs, LR and SR channels, users per cooperation cluster, and transmission rate. To evaluate our solution, we focus on a stadium topology to demonstrate the significant gains of optimized traffic offloading in conventional and D2D ultra dense wireless networks with/without cache-enabled devices. In addition, we propose a sub-optimal traffic offloading approach to provide real-time solutions in ultra dense D2D cooperative networks. A dynamic tree-based traffic offloading approach is proposed to assign users' connections consecutively based on a tree having BSs/APs and mobile terminals as nodes. Performance results and complexity analysis are presented to show that the proposed approach is able to provide near-optimal solutions with very low time complexity.

5. Formulating channel allocation to cluster heads as an optimization problem aiming at minimizing interference by minimizing the channel reuse and maximizing the distance between non-orthogonal channels. The solution of the channel allocation is based on solutions obtained from the optimal traffic offloading problem, where cluster heads, SR and LR users are determined. Due to non-orthogonality of the channels, user service target rate may not be achievable. We then present possible solutions for allocating connections for the users affected by interference either to other mobile terminals cluster heads or BSs/APs. Results showed reduction in outage percentage with a tradeoff cost of LR channel allocation. In addition, due to the high complexity of the optimal channel allocation, real-time and fast solutions are needed with traffic offloading in real time dense heterogeneous networks. Accordingly, we propose a sub-optimal approach for resource management in dense D2D cooperative networks where traffic offloading is simultaneously considered with channel allocation. We present a dynamic tree-based resource management where users are assigned consecutively based on a tree having BSs/APs and mobile terminals as nodes. The mobile terminals serving as cluster heads in the tree are assigned SR channels while performing traffic offloading.

7.2 Standardization Considerations

In this section, we highlight existing standards and ongoing standardization activities that will help facilitate the implementation of multiple wireless interfaces usage and cooperative common content distribution architectures in practical wireless networks.

7.2.1 Inter-Operation between Radio Access Technologies

The area of heterogeneous networks is gaining a lot of interest in the literature recently especially that applications are becoming more diverse. In addition, operators are looking for enhancing the quality of service and satisfying users expectations and quality of experience.

The idea behind HetNets is to overlay existing macro networks with additional infrastructure in the form of smaller low-power low-complexity access nodes. A mobile terminal can then take advantage of the co-existence of the available multiple wireless interfaces to achieve performance gains. This raises the issue of inter-operation between the various technologies. In 3GPP, there are several standardization efforts that deal with the coexistence of multiple RATs, e.g. [3, 120, 121]. The main focus related to the coexistence of multiple radio access technologies is on achieving seamless handover while selecting the best wireless interface.

In addition, LTE-Advanced standards are evolving towards supporting advanced features such as small cell deployments and LTE D2D communications. The vision towards 5G is to allow the dynamic utilization of spectrum and multiple access technologies for the best delivery of services, including D2D communications, spectrum refarming and radio access infrastructure sharing [4, 9].

Along the same direction, mobile devices manufacturers are working on introducing smart switching between cellular and WLAN. In the current mobile devices, the traffic is offloaded directly to WiFi when WLAN is available. Smart mode option allowing auto-switching between WiFi and cellular data networks are implemented in some smartphones such as iOS9 Iphones, Samsung Galaxy S5 and Sony Experia Z3. This allows the device to switch from WiFi to cellular network only when WiFi network is not stable with poor connectivity.

Beyond managing the switching between various RATs, standardization, operators and manufacturers efforts dealing with the joint and collaborative transmission and reception of a MT over multiple RATs are still lagging behind the research efforts in this direction.

7.2.2 Device-to-Device Communications

D2D cooperation is a solution to increase system coverage and capacity, and to reduce energy consumption of mobile devices for common content distribution.

In the cooperative network, mobile terminals can cooperate to receive data from other mobile terminals using a short range wireless technology (such as LTE-Direct, WiFi-Direct, Bluetooth or WiFi ad hoc mode). This is becoming feasible with the recent development of state-of-the-art wireless standards.

First, the Bluetooth has evolved to support advanced functionalities and higher bit rates for ad hoc communications among a group of MTs. The two most prevalent implementations of the specification are Bluetooth basic rate/enhanced data rate, which was adopted as version 2.0/2.1, and Bluetooth with low energy, which was adopted as version 4.0/4.1/4.2/5.0. Bluetooth v5.0 is the most recent version of the standard and it includes a special mode of operation characterized by significantly reduced energy consumption, increase the range, speed (up to 2Mbps) and broadcast messaging capacity of Bluetooth applications [122]. The main limitations of using Bluetooth include its relatively limited radio range (typically around 10 m) in addition to inter-system interference since it utilizes unlicensed spectrum.

Second, WiFi interface in ad hoc mode can be the most suitable alternative for Bluetooth to communicate with other MTs over SR links, taking into account the recent development of the WiFi-Direct specifications [123]. WiFi-Direct is a WiFi standard enabling devices to easily connect with each other without requiring a wireless access point, with a high bit rate (up to around 100 Mbps) over ranges that can reach up to 100 m.

Third, Qualcomm Incorporated introduced FlashLinQ as an advancement in peer-to-peer wireless technology. FlashLinQ uses a synchronous TDD OFDMA technology operating on dedicated licensed spectrum and provides high discovery range up to 1 km and capacity of thousands of nearby devices [124].

In addition, LTE-Direct is an innovative device-to-device technology to scale up from todays proximal discovery solutions. LTE-Direct was proposed as Release 12 3GPP feature [125]. It uses licensed spectrum framework where devices transmit at nominal mobile power levels. LTE-Direct provides a synchronous system where devices boradcast their services and wakeup periodically to discover all devices within range, in contrast to WiFi-Direct where discovery is based on two step asynchronous messages for device and service discovery.

The main challenge with D2D communications is to still to keep the interference to the primary cellular network at tolerable levels, especially when networks underlay the LTE-A cellular network. Mechanisms for D2D communications session setup and management are also under extensive search.

7.3 Open Research Directions

There are several open research problems that need to be addressed in order to implement optimized resource management and cooperative common content distribution architectures and algorithms in practical wireless networks. In this section, we highlight open research directions that complement the discussions presented throughout the dissertation.

7.3.1 Advanced Practical Protocols and Test Beds

First, there are no standardized protocols for network selection and traffic splitting in heterogeneous networks. The switching and splitting techniques proposed in the literature showed performance gains, however, thresholds for network switching and handover, traffic splitting ratios and strategies are not yet defined. In addition, traffic splitting and the use of multiple networks simultaneously is not yet supported even though our results showed that using WiFi and cellular links simultaneously may achieve higher throughput than when using one link with a slight cost of energy consumption.

Second, there are also no standardized content distribution protocols that enable the mobile terminals to connect intelligently to the cellular network or other mobile devices to download the content cooperatively. Accordingly, there is an urge to develop practical protocols to enable the mobile terminals with/without the assistance of the network central entity to decide on downloading data either from BS/AP via LR links or from another cluster head mobile terminal via SR, and on cooperating in data transmission to other mobile terminals. In addition, the traffic offloading objectives and user preferences determine whether to minimize the cellular cost, reduce the required cellular resources, minimize the energy consumption of the MTs or enhance QoS and QoE.

To achieve performance gains, future research investigation is needed to overcome several implementation challenges: (1) mobile devices manufacturers need to allow the simultaneous use of different wireless interfaces, (2) packet re-ordering is one of the major challenges when using traffic splitting, where selected data chunks are downloaded via both interfaces, (3) content distribution protocols in real-time networks need to accommodate for the fast network conditions variations including user mobility, (4) limited number of orthogonal non-overlapping channels which leads to interference and reduce system performance.

Existing work in the literature addressed optimized solutions, challenges, complexities whose outcome can be used in designing such protocols. Moreover, the existence of such intelligent cooperative protocols enables the implementation of practical test beds and actual implementation to achieve performance gains and serve towards 5G 2020 vision.

7.3.2 Practical Implementations and Incentives for Device-to-Device Communications

Content sharing through device-to-device communications has been proven to be a promising method to offload the traffic of base stations. If some user devices have cached a few popular on-demand contents, other interested neighbor mobile terminals can reuse these contents through D2D communications, in which the contents are directly transmitted to the mobile terminals from the content owners. Hereby, the base station would only transmit contents which are not locally available instead of transmitting the same popular contents multiple times. Therefore, the traffic of the BSs is significantly offloaded using D2D cooperation. This improves spectrum utilization, increases network throughput, and reduces average access delay for mobile terminals. However, in reality, users are selfish and only care about their own preferences. On the other hand, the base station aims at minimizing its traffic load and transmission cost by offloading to D2D communication. Accordingly, to motivate the users to participate in the cooperative content sharing, incentives for cooperation are needed. As a suggestion, the operators can reduce the cellular cost for the users acting as cluster heads, or offer them special subscriptions to encourage them to participate in content sharing [61–63].

In D2D cooperative networks, the mobile terminals sense the network parameters for nearby mobile and service discovery. Accordingly, signaling protocols are needed to intelligently sense the network while reducing the energy consumption and avoid any additional delays that may affect the overall performance gains and QoE. Furthermore, the overhead due to signaling needs to be assessed in practical D2D network, especially, in ultra dense networks where thousands of users are simultaneously connected.

Finally, the content reliability and availability becomes more challenging in dense networks where intelligent caching protocols are needed to guarantee that users is able to download all the data requested with the best performance. The mobility of the users mobile terminals as well as the security aspects in D2D cooperation are interesting research topics that need further investigation.

7.3.3 Resource Management in Ultra Dense Networks

Due to the growing number of mobile devices, network densification can also be a solution to meet user demands and expectations. An ultra dense network includes densely deployed small cells, macro base stations and large number of mobile terminals. UDNs face multiple significant challenges, including interference, resource management and cost.

In UDNs, inter-cell interference is caused by spectrum scarcity, where the spectrum resource is not able to cope with the increased demand. Frequency reuse techniques among different cells are then needed to support the large number of MTs. This will be further complicated by D2D dense deployments, close distance and irregular deployments. In addition, in UDNs, both the macro cells and small cells are cross deployed throughout the network, and small cells may reuse the same channel from the macro cell, which causes multi-tier interference. Accordingly, intelligent resource partitioning on frequency-domain, time-domain, and spatial-domain in either network side or mobile terminal side are needed to mitigate interference.

To achieve performance gains, several resource management issues are needed including energy management and spectrum management. The most energy consuming component in the cellular network is the base station. For this reason, a number of research efforts have been developed to propose different strategies to reduce the energy cost of the BS via power control and scheduling. For instance, in a small cell network, BSs are not equally used which leads to serious energy waste. Energy partitions can be then proposed, where the BSs associate with each other, taking turns in powered-on and powered-off states. Spectrum management is another aspect that will impact the system performance of UDN. Dynamic spectrum access protocols may be developed for enhanced spectrum sharing based on the primary/secondary strategy, where the secondary user can utilize the channel if the primary user is not using it. Accordingly, advanced resource management are needed to achieve the full potential gains of ultra dense networks.

List of Figures

List of Tables

Bibliography

[1] Cisco, "Cisco Visual Networking Index: Forecast and Methodology, 2016-2021," *White Paper, Cisco*, 2017.

[2] J. G. Andrews, S. Buzzi, W. Choi, S. Hanly, A. Lozano, A. C. K. Soong and J. C. Zhang, "What Will 5G Be?," *IEEE Journal on Selected Areas in Communications, Special Issue on 5G Communication Systems arXiv:1405.2957v1*, September 2014.

[3] 3GPP TR 22.934 version 11.0.0 Release 11, "Feasibility Study on 3GPP System to Wireless Local Area Network (WLAN) Interworking," 2012.

[4] 3GPP TR 36.932 version 12.1.0 Release 12, "LTE; Scenarios and requirements for small cell enhancements for E-UTRA and E-UTRAN," 2014.

[5] 3GPP TS 22.278 version 8.5.0 Release 8, "Service requirements for the Evolved Packet System (EPS)," 2008.

[6] 3GPP TR 36.839 version 11.0.0 Release 11, "Evolved Universal Terrestrial Radio Access (E-UTRA); Mobility enhancements in heterogeneous networks," 2012.

[7] IMT-2020 (5G) Promotion Group, "IMT Vision towards 2020 and Beyond," February 2014.

[8] IMT-2020 (5G) Promotion Group, "5G Vision and Requirements," 2014.

[9] 3GPP TR 36.843 version 0.2.0 Release 12, "Study on LTE Device to Device Proximity Services; Radio Aspects," 2013.

[10] E. Gustafsson and A. Jonsson, "Always Best Connected," *IEEE Wireless Communications*, vol. 10, no. 1, pp. 49–55, February 2003.

[11] Q. Nguyen-Vuong, N. Agoulmine and Y. Ghamri-Doudane, "A User-Centric and Context-Aware Solution to Interface Management and Access Network Selection in Heterogeneous Wireless Environments," *Computer Networks*, vol. 52, no. 18, pp. 3358–3372, December 2008.

[12] M.-R. Ra, J. Paek, A. B. Sharma, R. Govindan, M. H. Krieger and M. J. Neely, "Energy-delay Tradeoffs in Smartphone Applications," *in Proceedings of the Eight International Conference on Mobile Systems, Applications, and Services* , June 2010.

[13] N. Abbas and J. J. Saade, "A Fuzzy Logic Based Approach for Network Selection in WLAN/3G Heterogeneous Network," *in Proceedings of the IEEE Consumer Communications and Networking Conference*, January 2015.

[14] M. El Helou, M. Ibrahim, S. Lahoud, K. Khawam, D. Mezher and B. Cousin, "A Network-Assisted Approach for RAT Selection in Heterogeneous Cellular Networks," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 6, pp. 1055–1067, 2015.

[15] D. Ma and M. Ma, "Network Selection and Resource Allocation for Multicast in HetNets," *Journal of Network and Computer Applications*, vol. 43, no. , p. 1726, 2014.

[16] P. Naghavi, S. Hamed Rastegar, V. Shah-Mansouri and H. Kebriaei, "Learning RAT Selection Game in 5G Heterogeneous Networks," *IEEE Wireless Communications Letters*, vol. 5, no. 1, pp. 52–55, 2016.

[17] N. Hasan, W. Ejaz, N. Ejaz, H. S. Kim, A. Anpalagan and M. Jo, "Network Selection and Channel Allocation for Spectrum Sharing in 5G Heterogeneous Networks," *IEEE Access*, vol. PP, no. 99, pp. 1–11, 2016.

[18] K. Khawam, S. Lahoud, M. Ibrahim, M. Yassin, S. Martind, M. El Helou and F. Moetye, "Radio Access Technology Selection in Heterogeneous Networks," *Physical Communication*, vol. 18, no. , pp. 125–139, 2016.

[19] E. Patouni, N. Alonistioti and L. Merakos, "Modeling and Performance Evaluation of Reconfiguration Decision Making in Heterogeneous Radio Network Environments," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 4, pp. 1887–1900, 2010.

[20] X. Liu, X. Fang, X. Chen and X. Peng, "A Bidding Model and Cooperative Game-Based Vertical Handoff Decision Algorithm," *Journal of Network and Computer Applications*, vol. 34, no. , p. 12631271, 2011.

[21] R. Trestian, Member, O. Ormond and G.-M. Muntean, "Energy-Quality-Cost Tradeoff in a Multimedia-Based Heterogeneous Wireless Network Environment," *IEEE Transactions on Broadcasting*, vol. 59, no. 2, pp. 340–357, 2013.

[22] S. Singh and J.G. Andrews, "Joint Resource Partitioning and Offloading in Heterogeneous Cellular Networks," *IEEE Transactions on Wireless Communications*, vol. 13, no. 2, pp. 888–901, 2014.

[23] S.-L. Tsao, Ch.-W. Wang, Y.-C. Lin and R.-G. Cheng, "A Dynamic Load-Balancing Scheme for Heterogeneous Wireless Networks," *in Proceedings of the IEEE Wireless Communications and Networking Conference*, April 2014.

[24] S. Cai, L. Duan, J. Wang, S. Zhou and R. Zhang, "Incentive Mechanism Design for Delayed WiFi Offloading," *in Proceedings of the IEEE International Conference on Communications*, June 2015.

[25] W.S. Lai, T.H. Chang and T.-S. Lee, "Joint Power and Admission Control for Spectral and Energy Efficiency Maximization in Heterogeneous OFDMA Networks," *IEEE Transactions on Wireless Communications*, vol. PP, no. 99, pp. 1–16, 2016.

[26] J.-O. Kim, T. Ueda and S. Obana, "MAC-Level Measurement Based Traffic Distribution over IEEE 802.11 Multi-Radio Networks," *IEEE Transaction on Consumer Electronics*, vol. 54, no. 3, pp. 1185–1191, August 2008.

[27] J.-O. Kim, "Feedback-Based Traffic Splitting for Wireless Terminals with Multi-Radio Devices," *IEEE Transaction on Consumer Electronics*, vol. 56, no. 2, pp. 476–482, May 2010.

[28] R. Yang, Y. Chang, J. Sun and D. Yang, "Traffic Split Scheme Based on Common Radio Resource Management in an Integrated LTE and HSDPA Networks," *in Proceedings of Vehicular Technology Conference*, September 2012.

[29] X. Gelabert, O. Sallent, J. Perez-Romero and R. Agusti, "Performance Evaluation of Radio Access Selection Strategies in Constrained Multi-Access/Multi-Service Wireless Networks," *Computer Networks*, vol. 55, no. 1, pp. 173–192, January 2011.

[30] J. Luo, R. Mukerjee, M. Dillinger, E. Mohyeldin and E. Schulz, "Investigation of Radio Resource Scheduling in WLANs Coupled with 3G Cellular Network," *IEEE Communications Magazine*, vol. 41, no. 6, pp. 108–115, June 2003.

[31] H. Lian, X. Yan, L. Weng, Z. Feng, Q. Zhang and P. Zhang, "Efficient Traffic Allocation Scheme for Multi-flow Distribution in Heterogeneous Networks," *in Proceedings of IEEE Globecom Workshops*, December 2013.

[32] S. Dimatteo, P. Hui, B. Han and V.O.K Li, "Cellular Traffic Offloading Through WiFi Networks," *in Proceedings of the 8th IEEE International Conference on Mobile Adhoc and Sensor Systems (MASS)*, October 2011.

[33] J. Stadler and G. Pospischil, "Simultaneous Usage of WLAN and UTRAN for Improved Multimedia and Data Applications," *in Proceedings of the 11th International Telecommunications Network Strategy and Planning Symposium*, June 2004.

[34] S.-N. Yang, Sh.-W. Ho, Y.-B. Lin and C.-H. Gan, "A Multi-RAT Bandwidth Aggregation Mechanism with Software-Defined Networking," *Journal of Network and Computer Applications*, vol. 61, no. , p. 189198, 2016.

[35] H. Ju, B. Liang, J. Li, Y. Long and X. Yang, "Adaptive Cross-Network Cross-Layer Design in Heterogeneous Wireless Networks," *IEEE Transactions on Wireless Communications*, vol. 14, no. 2, pp. 655–669, 2015.

[36] J. Li, J. Zheng, Q. Liu and X. Yang, "Delay Performance Optimization of Multiaccess for Uplink in Heterogeneous Networks," *in Proceedings of the 79th IEEE Vehicular Technology Conference*, May 2014.

[37] Y. Song, Y. Han and Y. Choi, "Radio Resource Management Based on QoE-Aware Model for Uplink Multi-Radio Access in Heterogeneous Networks," *in Proceedings of the 79th IEEE Vehicular Technology Conference*, May 2014.

[38] Y. Zhou and W. Zhuang, "Throughput Analysis of Cooperative Communication in Wireless Ad Hoc Networks With Frequency Reuse," *IEEE Transactions on Wireless Communications*, vol. 14, no. 1, pp. 205–218, 2015.

[39] S. M. A. Kazmi, N. H. Tran, W. Saad, Z. Han, T. M. Ho, T. Z. Oo and C. S. Hong, "Mode Selection and Resource Allocation in Device-to-Device Communications: A Matching Game Approach," *IEEE Transactions on Mobile Computing*, vol. PP, no. 99, pp. 1–1, 2017.

[40] H. Bagheri and M. Katz, "A Resource Allocation Mechanism for Enhancing Spectral Efficiency and Throughput of Multi-link D2D Communications," *in Proceedings of the 25th Annual International Symposium on Personal, Indoor, and Mobile Radio Communication*, September 2014.

[41] M. Ahmad, M. Naeem, A. Ahmed, M. Iqbal, A. Anpalagan and W. Ejaz, "Utility Based Resource Management in D2D Networks Using Mesh Adaptive Direct Search Method," *in Proceedings of the IEEE 84th Vehicular Technology Conference* , September 2016.

[42] E. Yaacoub, F. Filali and A. Abu-Dayya, "QoE Enhancement of SVC Video Streaming Over Vehicular Networks Using Cooperative LTE/802.11p Communications," *IEEE Journal in Selected Topics in Signal Processing*, vol. 9, no. 1, pp. 37–49, 2015.

[43] K. Xie, X. Wang, X. Liu, J. Wen and J. Cao, "Interference-Aware Cooperative Communication in Multi-radio Multi-channel Wireless Networks," *IEEE Transactions on Computers*, vol. PP, no. 99, pp. 1–14, 2015.

[44] Y. Li, L. Zhang, X. Tan and B. Cao, "An Advanced Spectrum Allocation Algorithm for the Across-Cell D2D Communication in LTE Network with Higher Throughput," *China Communications*, vol. 13, no. 4, pp. 30–37, 2016.

[45] Y.-C. Lai, P.Lin, W. Liao and C.-M.Chen, "A Region-Based Clustering Mechanism for Channel Access in Vehicular Ad Hoc Networks," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 1, pp. 83–93, 2011.

[46] K.-J. Peng and Z. Tsai, "Distortion and Cost Controlled Video Streaming in a Heterogeneous Wireless Network Environment," *in Proceedings of the 17th International Symposium on Personal, Indoor and Mobile Radio Communications*, September 2006.

[47] M.-F. Leung, and S.-H. G. Chan, "Broadcast-Based Peer-to-Peer Collaborative Video Streaming Among Mobiles," *IEEE Transactions on Broadcasting*, vol. 53, no. 1, pp. 350–361, 2007.

[48] C. Korn, D. S. Michalopoulos and R. Schober, "Power Saving Potential of Cooperative Communication: A Two-Dimensional Approach," *in Proceedings of the 10th International Conference on on Systems, Communications and Coding*, February 2015.

[49] W. Xie, B. Zhou, E. Liu and T. Zhou, "Research on the Computing Network Topology Distribution Based on Energy Consumption for Cooperative Communication System," *in Proceedings of the 7th International Symposium on Computational Intelligence and Design*, December 2014.

[50] T. Han, Y. Feng, J. Wang, L. Wang, Q. Li and Y. Han, "On Energy Efficiency of the Nearest-Neighbor Cooperative Communication in Heterogeneous Networks," *in Proceedings of the IEEE International Conference on Communication*, June 2015.

[51] E. Yaacoub, L. Al-Kanj, Z. Dawy, S. Sharafeddine, F. Filali and A. Abu-Dayya, "A Nash Bargaining Solution for Energy-Efficient Content Distribution over Wireless Networks with Mobile-to-Mobile Cooperation," *in Proceedings of the 4th International Wireless and Mobile Networking Conference* , October 2011.

[52] L. Al-Kanj, M. Abdallah and Z. Dawy, "On Optimal Mobile Terminal Grouping in Energy Aware Cooperative Content Distribution Networks,"

in Proceedings of the International Wireless and Mobile Networking Conference, April 2012.

[53] L. Al-Kanj, L. Saad and Z. Dawy, "A Game Theoretic Approach for Content Distribution over Wireless Networks with Mobile-to-Mobile Cooperation," *in Proceedings of the 22nd International Symposium on Personal Indoor and Mobile Radio Communications*, September 2011.

[54] L. Al-Kanj, H. V. Poor and Z. Dawy, "Optimal Cellular Offloading via Device-to-Device Communication Networks With Fairness Constraints," *IEEE Transactions on Wireless Communications*, vol. 13, no. 8, pp. 4628–4643, 2014.

[55] L. Al-Kanj and Z. Dawy, "Offloading Wireless Cellular Networks via Energy-Constrained Local Ad Hoc Networks," *in Proceedings of the IEEE Global Telecommunications Conference* , December 2011.

[56] R. Dinga, C. Hava Munteanb and G.M. Munteana, "Energy-Efficient Device-Differentiated Cooperative Adaptive Multimedia Delivery Solution in Wireless Networks," *Journal of Network and Computer Applications*, vol. 58, no. , pp. 194–207, 2015.

[57] H. Xiao and S. Ouyang, "Power Control Game in Multisource Multirelay Cooperative Communication Systems With a Quality-of-Service Constraint," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 1, pp. 41–50, 2015.

[58] Z. Mo, W. Su, S. Batalama and J. D. Matyjas, "Cooperative Communication Protocol Designs Based on Optimum Power and Time Allocation," *IEEE Transactions on Wireless Communications*, vol. 13, no. 8, pp. 4283–4296, 2014.

[59] W. Shi, G. Zhao and Z. Chen, "Relay Selection and Power Control for Energy-Efficient Cooperative Multicast Communication," *in Proceedings of the 81st IEEE International Vehicular Technology Conference*, May 2015.

[60] P. Li, S. Guo and J. Hu, "Energy-Efficient Cooperative Communications for Multimedia Applications in Multi-Channel Wireless Networks," *IEEE Transactions on Computers*, vol. 64, no. 6, pp. 1670–1679, 2015.

[61] H. Zhu, Y. Cao, B. Liu and T. Jiang, "Energy-Aware Incentive Mechanism for Content Sharing Through Device-to-Device Communications," *in Proceedings of the IEEE Global Communications Conference*, December 2016.

[62] F. Alotaibi, S. Hosny, H. El Gamal and A. Eryilmaz, "A Game Theoretic Approach to Content Trading in Proactive Wireless Networks," *IEEE International Symposium on Information Theory*, 2015.

[63] Z. Chen, Y. Liu, B. Zhou and M. Tao, “Caching Incentive Design in Wireless D2D Networks: A Stackelberg Game Approach,” *in Proceedings of the IEEE International Conference on Communications*, May 2016.

[64] A. Afzal, S. A. R. Zaidi, D. McLernon and M. Ghogho, “On the Analysis of Cellular Networks with Caching and Coordinated Device-to-Device Communication,” *in Proceedings of the IEEE International Conference on Communications*, May 2016.

[65] Y. Sun, Z. Chen and H. Liu, “Delay Analysis and Optimization in Cache-Enabled Multi-Cell Cooperative Networks,” *in Proceedings of the IEEE Global Communications Conference*, December 2016.

[66] Y. Long, Y. Cai, D. Wu and L. Qiao, “Content-Related Energy Efficiency Analysis in Cache-Enabled Device-to-Device Network,” *in Proceedings of the 8th International Conference on Wireless Communications and Signal Processing*, October 2016.

[67] B. Chen, C. Yang and Z. Xiong, “Optimal Caching and Scheduling for Cache-Enabled D2D Communication,” *IEEE Communications Letters*, 2017.

[68] Y. Wang, Y. Li, W. Wang and M. Song, “A Locality-based Mobile Caching Policy for D2D-based Content Sharing Network,” *in Proceedings of the IEEE Global Communications Conference* , December 2016.

[69] B. Chen and C. Yang, “Energy Costs for Traffic Offloading by Cache-enabled D2D Communications,” *in Proceedings of the IEEE Wireless Communications and Networking Conference*, April 2016.

[70] W. Wang, R. Lan, J. Gu, A. Huang, H. Shan and Z. Zhang, “Edge Caching at Base Stations with Device-to-Device Offloading,” *IEEE Access*, vol. PP, no. 99, pp. 1–1, 2017.

[71] B. Chen; C. Y. Yang and G. Wang, “High Throughput Opportunistic Cooperative Device-to-Device Communications With Caching,” *IEEE Transactions on Vehicular Technology*, vol. PP, no. 99, pp. 1–1, 2017.

[72] A. Masucci, S. E. Elayoubi and B. Sayra, “Flow level analysis of the offloading capacity of D2D communications,” *in Proceedings of the IEEE Wireless Communications and Networking Conference*, April 2016.

[73] M. Deressa Amentie, M. Sheng, J. Song and J. Liu, “Minimum Delay Guaranteed Cooperative Device-to-Device Caching in 5G Wireless Networks,” *in Proceedings of the 8th International Conference on Wireless Communications and Signal Processing*, October 2016.

[74] B. Chen, C. Yang and G. Wang, "Cooperative Device-to-Device Communications with Caching," *in Proceedings of the IEEE 83rd Vehicular Technology Conference*, May 2016.

[75] G. S. Park, W. Kim, S. H. Jeong and H. Song, "Smart Base Station-Assisted Partial-Flow Device-to-Device Offloading System for Video Streaming Services," *IEEE Transactions on Mobile Computing*, vol. PP, no. 99, pp. 1–1, 2016.

[76] X. Zhang, Y. Wang, R. Sun and D. Wang, "Clustered Device-to-Device Caching Based on File Preferences," *in Proceedings of the IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications*, September 2016.

[77] J. Jiang, S. Zhang, B. Li and B. Li, "Maximized Cellular Traffic Offloading via Device-to-Device Content Sharing," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 1, pp. 82–91, 2016.

[78] M. Kamel, W. Hamouda and A. Youssef, "Ultra-Dense Networks: A Survey," *IEEE Communications Surveys and Tutorials*, vol. 18, no. 4, pp. 2522–2545, 2016.

[79] Q. Ren, J. Fan, X. Luo, Z. Xu and Y. Chen, "Analysis of Spectral and Energy Efficiency in Ultra-Dense Network," *in Proceedings of the IEEE International Conference on Communication Workshop*, June 2015.

[80] S. Samarakoon, M. Bennis, W. Saad, M. Debbah and M. Latva-aho, "Ultra Dense Small Cell Networks: Turning Density Into Energy Efficiency," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 5, pp. 1267–1280, 2016.

[81] C. Niu, Y. Li, R. Q. Hu and F. Ye, "Fast and Efficient Radio Resource Allocation in Dynamic Ultra-Dense Heterogeneous Networks," *IEEE Access*, vol. 5, no. , pp. 1911–1924, 2017.

[82] Q. Meng, S. Zhong and Y. Zou, "Offloading in Dynamic Ultra Dense Small Cell Networks," *in Proceedings of the 7th IEEE International Conference on Software Engineering and Service Science* , August 2016.

[83] V. Jevremovic, "How to design better wireless networks for stadiums," *Webinar and Case Study*, September 2015.

[84] N. Morozs, T. Clarke and D. Grace, "Distributed Heuristically Accelerated Q-Learning for Robust Cognitive Spectrum Management in LTE Cellular Systems," *IEEE Transactions on Mobile Computing*, vol. 15, no. 4, pp. 817–825, 2016.

[85] N. Morozs, T. Clarke and D. Grace, "Intelligent Dynamic Spectrum Access in Cellular Systems with Asymmetric Topologies and Non-Uniform Traffic Loads," *in Proceedings of the 82nd Vehicular Technology Conference*, September 2015.

[86] M. N. Islam, S. Subramanian, A. Partyka and A. Sampath, "Coverage and Capacity of 28 GHz Band in Indoor Stadiums," *in Proceedings of the IEEE Wireless Communications and Networking Conference*, April 2016.

[87] A. . Kaya, D. Calin and H. Viswanathan, "On the Performance of Stadium High Density Carrier Wi-Fi Enabled LTE Small Cell Deployments," *in Proceedings of the IEEE Wireless Communications and Networking Conference*, March 2015.

[88] K. Yang, Z. Luo, H. Zhuang, J. Zhang and Q. Gao, "Flexible Carrier Utilization in Dense Stadium," *in Proceedings of the IEEE 84th Vehicular Technology Conference*, September 2016.

[89] N. Morozs, T. Clarke, D. Grace and Q. Zhao, "Distributed Q-Learning Based Dynamic Spectrum Management in Cognitive Cellular Systems: Choosing the Right Learning Rate," *IEEE Symposium on Computers and Communications*, 2014.

[90] S. Moballegh and B. Sirkeci, "Analysis of Cooperative Communication in One-Dimensional Dense Ad-Hoc Networks," *in Proceedings of the IEEE International Conference on Communication*, June 2015.

[91] S. Krishnasamy and S. Shakkottai, "Spectrum Sharing and Scheduling in D2D-Enabled Dense Cellular Networks," *in Proceedings of the 13th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks*, May 2015.

[92] I. AlQerm and B. Shihada, "A Cooperative Online Learning Scheme for Resource Allocation in 5G Systems," *in Proceedings of the IEEE International Conference on Communication*, May 2016.

[93] S. Iskounen, T. M. T. Nguyen, S. Monnet and L. Hamidouche, "Device-to-Device Communications Using Wi-Fi Direct for Dense Wireless Networks," *in Proceedings of the 7th International Conference on the Network of the Future*, November 2016.

[94] M. A. Kader, E. Bastug, M. Bennis, E. Zeydan, A. Karatepe, A. S. Er and M. Debbah, "Leveraging Big Data Analytics for Cache-Enabled Wireless Networks," *in Proceedings of the IEEE Globecom Workshops*, December 2015.

[95] X. Song, Y. Geng, X. Meng, J. Liu, W. Lei and Y. Wen, “Cache-Enabled Device to Device Networks with Contention Based Multimedia Delivery,” *IEEE Access*, vol. 5, no. , p. , 2017.

[96] J. Han, M. Kamber and J. Pei, *Data Mining Concepts and Techniques.* Morgan Kaufmann, 2012.

[97] J. Kellokoski, J. Koskinen and T. Hamalainen, “Power Consumption Analysis of the Always-Best-Connected User Equipment,” *in Proceedings of the 5th International Conference on New Technologies, Mobility and Security*, May 2012.

[98] RapidMiner Manual, “RapidMiner 5.0 Manual,” *Rapid-I GmbH*, 2010.

[99] N. Abbas, Z. Dawy, H. Hajj and S. Sharafeddine, “Energy-Throughput Tradeoffs in Cellular/WiFi Heterogeneous Networks with Traffic Splitting,” *in Proceedings of the IEEE Wireless Communications and Networking Conference*, April 2014.

[100] D. Bethanabhotla, G. Caire and M. J. Neely, “Adaptive Video Streaming for Wireless Networks with Multiple Users and Helpers,” *arXiv preprint arXiv:1304.8083v2*, April 2014.

[101] M. J. Neely, *Stochastic Network Optimization with Application to Communication and Queueing Systems.* Morgan and Claypool, 2010.

[102] [ITU-T P.10/G.100] Recommendation ITU-T P.10/G.100 (2016), “Amendment 1: New Appendix I Definition of Quality of Experience (QoE),” 2006.

[103] N. Rjaibi, L. Ben Arfa Rabai and M.Limam, “Modeling the Prediction of Student’s Satisfaction in Face to Face Learning: An Empirical Investigation,” *in Proceedings of the International Conference on Education and e-Learning Innovations*, July 2012.

[104] M. H. Pinson and S. Wolf, “A New Standardized Method for Objectively Measuring Video Quality,” *IEEE Transactions on Broadcasting*, vol. 50, no. 3, pp. 312–322, 2004.

[105] K.Kilkki, “Quality of Experience in Communications Ecosystem,” *Journal of Universal Computer Science*, vol. 14, no. 5, pp. 615–624, March 2008.

[106] [ITU-T P.1202] Recommendation ITU-T P.1202 (2012), “Parametric Non-Intrusive Assessment of Audiovisual Media Streaming Quality,” October 2012.

[107] European Telecommunications Standards Institute, ETSI TR 102 714 V1.1.1, “Speech and Multimedia Transmission Quality (STQ); Multimedia Quality Measurement; End-to-End Quality Measurement Framework,” 2011.

[108] [ITU-T J.340] Recommendation ITU-T J.340 (2010), “Reference Algorithm for Computing Peak Signal to Noise Ratio of a Processed Video Sequence with Compensation for Constant Spatial Shifts, Constant Temporal Shift, and Constant Luminance Gain and Offset,” June 2010.

[109] Z. Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, “Image Quality Assessment: From Error Visibility to Structural Similarity,” *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[110] P. Szilagyi and C. Vulkan, “Network Side Lightweight and Scalable YouTube QoE Estimation,” *in Proceedings of the IEEE International Conference on Communications*, June 2015.

[111] [ITU-T P.1201] Recommendation ITU-T P.1201 (2013), “Parametric Non-Intrusive Assessment of Audiovisual Media Streaming Quality - Amendment 2: New Appendix III Use of ITU-T P.1201 for Non-Adaptive, Progressive Download Type Media Streaming,” December 2013.

[112] R. K. P. Mok, E. W. W. Chan and R. K. C. Chang, “Measuring the Quality of Experience of HTTP Video Streaming,” *in Proceedings of the 12th IFIP/ IEEE International Symposium on Integrated Network Management*, 2011.

[113] A. Khan, L. Sun, E. Jammeh and E. Ifeachor, “Quality of Experience-Driven Adaptation Scheme for Video Applications Over Wireless Networks,” *IET Communications In Special Issue on Video Communications over Wireless Networks*, vol. 4, no. 11, pp. 1337–1347, July 2010.

[114] A. Goldsmith, *Wireless Communications*. Cambridge University Press, 2005.

[115] G. Chartrand, F. Harary, and B. Q. Yue, “On the Out-Domination and in-Domination Numbers of a Digraph,” *Discrete Mathematics* , vol. 197198, pp. 179 – 183, 1999.

[116] G. Chartrand, P. Dankelmann, M. Schultz and H. C. Swart, “Twin Domination in Digraphs,” *Ars Comb.*, 2003.

[117] M. Bay and Y. Crama, “Introduction to AIMMS,” November 2004.

[118] British Association of Seating Equipment Suppliers, “Recommendations for the Specification and Use of Telescopic and Other Spectator Seating,” 2008.

[119] Cisco, "Channel Planning Best Practices," *White Paper, Cisco*, 2016.

[120] 3GPP TS 37.113 version 11.1.0 Release 11, "Digital cellular telecommunications system (Phase 2+); Universal Mobile Telecommunications System (UMTS); LTE; E-UTRA, UTRA and GSM/EDGE; Multi-Standard Radio (MSR) Base Station (BS) Electromagnetic Compatibility (EMC)," 2012.

[121] 3GPP TR 37.900 version 10.0.0 Release 10, "Digital cellular telecommunications system (Phase 2+); Universal Mobile Telecommunications System (UMTS); LTE; Radio Frequency (RF) requirements for Multicarrier and Multiple Radio Access Technology (Multi-RAT) Base Station (BS)," 2011.

[122] Specification of the Bluetooth System, "Bluetooth Core Specifications version: 5.0," December 2016.

[123] WiFi Alliance, "Wi-Fi Peer-to-Peer (P2P) Technical Specification Version 1.5," 2014.

[124] X. Wu, S. Tavildar, S. Shakkottai, T. Richardson, J. Li, R. Laroia and A. Jovicic, "FlashLinQ: A Synchronous Distributed Scheduler for Peer-to-Peer Ad Hoc Networks," *IEEE/ACM Transactions on Networking*, vol. 21, no. 4, pp. 1215–1228, 2013.

[125] Sajith Balraj Qualcomm Research, "LTE Direct Overview," 2012.