UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

Departament de Teoria
del Senyal i Comunicacions

Centre
Tecnològic
de Telecomunicacions
de Catalunya

Radio resource management techniques for QoS provision in 5G networks

PhD thesis dissertation

by

Eftychia Datsika

Submitted to the Universitat Politècnica de Catalunya (UPC)
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY

Advisors:

Christos Verikoukis, Ph. D.
Senior Researcher
Head of SMARTECH Department
Telecommunications Technological Center of Catalonia (CTTC/CERCA)

Angelos Antonopoulos, Ph. D.
Senior Researcher
SMARTECH Department
Telecommunications Technological Center of Catalonia (CTTC/CERCA)

PhD program on Signal Theory and Communications

Barcelona, September 2018

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Radio resource management techniques for QoS provision in 5G networks

by

Eftychia Datsika

This work is dedicated to Filippos-Michail Datsikas.

Abstract

As numerous mobile applications and over-the-top (OTT) services emerge and mobile Internet connectivity becomes ubiquitous, the provision of high quality of service (QoS) is more challenging for mobile network operators (MNOs). Research efforts focus on the development of innovative resource management techniques and have introduced the long term evolution advanced (LTE-A) communication standard. Novel business models make the growth of network capacity sustainable by enabling MNOs to combine their resources. The fifth generation (5G) mobile networks will involve technologies and business stakeholders with different capabilities and demands that may affect the QoS provision, requiring efficient radio resource sharing.

The need for higher network capacity has introduced novel technologies that improve resource allocation efficiency. Direct connectivity among user equipment terminals (UEs) circumventing the LTE-A infrastructure alleviates the network overload. Part of mobile traffic is offloaded to outband device-to-device (D2D) connections (in unlicensed spectrum) enabling data exchange between UEs directly or via UEs-relays. Still, MNOs need additional spectrum resources and infrastructure. The inter-operator network sharing concept has emerged motivating the adoption of virtualization that enables network slicing, i.e., dynamic separation of resources in virtual slices (VSs). VSs are managed in isolation by different tenants using software defined networking and encompass core and radio access network resources allocated periodically to UEs. When UEs access OTT applications, flows with different QoS demands and priorities determined by OTT service providers (OSPs) are generated. OSPs' policies should be considered in VS allocation. The coexisting technologies, business models and stakeholders require sophisticated radio resource management (RRM) techniques.

To that end, RRM is performed in a complex ecosystem. When D2D communication involves data concurrently downloaded by the mobile network, QoS may be affected by LTE-A network parameters (resource scheduling policy, downlink channel conditions). It is also affected by the relay selection, as UEs may not be willing to help unknown UE pairs and UEs' social ties in mobile applications may influence willingness for D2D cooperation. Thus, effective medium access control (MAC) mechanisms should coordinate D2D transmissions employing advanced techniques, e.g., network coding (NC). When UEs access OTT applications, OSPs' policies are not considered by MNOs in RRM and OSPs cannot apply flow prioritization. Network neutrality issues also arise when OSPs claim resources from MNOs aiming to minimize grade of service (GoS). OSPs' intervention may

delay flows' accommodation due to the time required for OSP-MNO interaction and the time the flows spent waiting for resources.

This thesis proposes novel solutions to the RRM issues of outband D2D communication and VS allocation for OSPs in 5G networks. We present a cooperative D2D MAC protocol that leverages the opportunities for NC in D2D communication under the influence of LTE-A network parameters and its throughput performance analysis. The protocol improves D2D throughput and energy efficiency, especially for UEs with better downlink channel conditions. We next introduce social awareness in D2D MAC design and present a social-aware cooperative D2D MAC protocol that employs UEs' social ties to promote the use of friendly relays reducing the total energy consumption. Motivated by the lack of approaches for OSP-oriented RRM, we present a novel flow prioritization algorithm based on matching theory that applies OSPs' policies respecting the network neutrality and the analysis of its GoS and delay performance. The algorithm maintains low overhead and delay without affecting fairness among OSPs. Our techniques highlight the QoS improvement induced by the joint consideration of different technologies and business stakeholders in RRM design.

Resumen

A medida que varias aplicaciones móviles y servicios over-the-top (OTT) surgen y el Internet móvil se vuelve ubicua, la prestación de alta calidad de servicio (QoS) es desafiante para los operadores de red móvil (MNOs). Los estudios de investigación se enfocan en técnicas innovadoras para la gestión de recursos de red y han resultado en la especificación del estándar de comunicación long term evolution advanced (LTE-A). Modelos comerciales nuevos hacen que el crecimiento de la capacidad de red sea sostenible al permitir que MNOs combinen sus recursos. La quinta generación (5G) de redes móviles implicará tecnologías y partes comerciales interesadas con varias habilidades y demandas que pueden afectar la provisión de QoS y demandan la gestión eficaz de recursos de radio.

La necesidad de capacidad de red más alta ha introducido tecnologías que hacen más eficiente la asignación de recursos. La conectividad directa entre terminales de equipos de usuarios (UEs) eludiendo la infraestructura LTE-A alivia la sobrecarga de red. Parte del tráfico es dirigido a conexiones de dispositivo a dispositivo (D2D) outband permitiendo la comunicación de UEs directamente o con relés. Los MNOs necesitan nuevos recursos de espectro e infraestructura. El intercambio de recursos entre MNOs ha surgido motivando la adopción de virtualización que realiza la segmentación de red i.e., la separación dinámica de recursos en trozos virtuales (VSs). Los VSs son administrados de forma aislada por inquilinos diferentes con software defined networking y abarcan recursos de red core y radio access asignadas periódicamente a UEs. Cuando UEs usan aplicaciones OTT, flujos de aplicación con demandas y prioridades definidas por proveedores de servicios OTT (OSPs) se generan. Las políticas de OSPs deben ser integradas en la asignación de VSs. La coexistencia de varias tecnologías y partes comerciales demanda técnicas sofisticadas de gestión de recursos radio (RRM).

Con ese fin, la RRM se realiza en un ecosistema complejo. Si la comunicación D2D involucra datos descargados simultáneamente por la red móvil, los parámetros de red LTE-A (política de scheduling de recursos, condiciones de canal downlink) afectan el QoS. La selección de relés afecta el rendimiento porque los UEs no desean siempre ayudar a UEs desconocidos. Las relaciones sociales de los UEs en aplicaciones móviles pueden determinar la voluntad para la comunicación cooperativa D2D. Por lo tanto, mecanismos de control de acceso al medio (MAC) deben coordinar las transmisiones D2D con técnicas avanzadas ej., codificación de red. Si los UEs usan servicios OTT, las políticas de OSPs no son consideradas en RRM y los OSPs no emplean flujos prioritarios. Problemas de neutralidad de red surgen cuando los OSPs reclaman recursos de MNOs para minimizar

el grado de servicio (GoS). La intervención de OSPs puede causar retraso en el servicio de flujos debido a la interacción OSP-MNO y el tiempo requerido para que los flujos reciban recursos.

Esta tesis presenta soluciones nuevas para los problemas RRM de comunicación D2D outband y asignación de VSs a OSPs en redes 5G. Proponemos un protocolo D2D MAC cooperativo que explota las oportunidades de NC bajo la influencia de parámetros de red LTE-A y su análisis de rendimiento. El protocolo mejora el rendimiento y la eficiencia energética especialmente para UEs con mejores condiciones de canal downlink. Introducimos la conciencia social en el D2D MAC y proponemos un protocolo que utiliza relaciones sociales de UEs para elegir relés-amigos y reduce el consumo de energía. Dada la falta de técnicas que aborden el problema RRM de OSPs presentamos un algoritmo que aplique políticas de OSPs y respete la neutralidad usando la teoría de matching, y su análisis de GoS y retraso. El algoritmo induce bajo coste y retraso sin afectar la imparcialidad entre OSPs. Estas técnicas demuestran la mejora de QoS gracias a la consideración de tecnologías y partes comerciales diferentes en RRM.

Acknowledgements

I am sincerely grateful to my thesis supervisors, Dr. Christos Verikoukis, who offered me the opportunity to begin this journey of research and the continuous guidance and valuable support to complete it successfully, and Dr. Angelos Antonopoulos, who generously provided constructive advice that helped me improve significantly my research skills. Thank you both for this fulfilling collaboration!

I would like to thank my tutor from UPC, Dr. Luis Alonso, for his assistance during my PhD studies. My thanks extend to Dr. Iordanis Koutsopoulos and Dr. Vasileios Vasilakis that agreed to review my work and their feedback has contributed to the improvement of the thesis. I would also like to thank my co-authors in the publications elaborated during this thesis for the fruitful collaboration and inspiring input.

I would like to express my gratitude to my parents and immediate family for their unconditional love and support during my years of studies and for having stood by me patiently in each of my endeavors. I also acknowledge the encouragement, insight and positive influence of Nikos. A very special thanks goes to him, as he has been beside me in every step of this effort. I am particularly indebted to my dearest friends Evaggelia, Katerina, Labrini and Maria (in alphabetical order) for their understanding and moral support during all these years.

Last, I would like to wish to all the people I collaborated with in Barcelona, Aveiro, Norrköping and Athens the best of luck and success in their future plans.

Contents

List of Figures

List of Tables

List of Acronyms

3GPP	3rd generation partnership project
5G	5th generation
ACK	Acknowledgment
ACNC-MAC	Adaptive cooperative network coding based medium access control
AP	Access point
API	Application programming interface
CN	Core Network
CQI	Channel quality indicator
CSMA/CA	Carrier sense multiple access with collision avoidance
CTMC	Continuous time Markov chain
D2D	Device-to-device
DCF	Distributed coordination function
DIFS	DCF interframe space
eNB	Evolved NodeB base station
EPC	Evolved packet core
ETC	Eager-to-cooperate
E-UTRAN	Evolved UMTS terrestrial radio access network
GoS	Grade-of-service
LTE	Long term evolution
LTE-A	Long term evolution advanced
MAC	Medium access control
MCS	Modulation and coding scheme
MNO	Mobile network operator
MTFP	Matching theoretic flow prioritization
MVNO	Mobile virtual network operator
NC	Network coding
OFDM	Orthogonal frequency division multiplexing
OSP	Over-the-top service provider
OTT	Over-the-top
PER	Packet Error Rate
QoS	Quality of service
RAN	Radio access network

RB	Resource block
RF	Radio frequency
RFC	Request-for-cooperation
SCD2D-MAC	Social aware cooperative D2D MAC
SCR	Social-cooperation-request
SDN	Software defined networking
SIFS	Short interframe space
SNR	Signal to noise ratio
SoA	State of the art
TBS	Transport block size
TTI	Transmission Time Interval
UE	User equipment terminal
UMTS	Universal mobile telecommunications system
VS	Virtual slice
Wi-Fi	Wireless fidelity
WNV	Wireless network virtualization

Chapter 1

Introduction

1.1 Overview and motivation

In the last few decades, the mobile communication has been established as a commodity available to the majority of the world's population and the technological advances in telecommunications have been widely integrated in modern network deployments. From the analog mobile radio systems of the '80s that only supported voice calls to the fourth generation (4G) of mobile networks that supports high-quality voice and data transportation, the mobile cellular networks have greatly evolved thanks to the joint research efforts of industry and academia. Nowadays, thanks to the development of the long term evolution advanced (LTE-A) standard, the 4G mobile networks can offer a variety of Internet-based services with high quality to the end users, which communicate via their user equipment terminals (UEs) using a plethora of mobile applications, either offered by the mobile network operators or other service providers that introduce over-the-top (OTT) applications accessed via Internet connections over mobile networks [4].

The escalating demands of mobile applications for network capacity have shaped the design of the upcoming fifth generation (5G) of mobile networks, which is expected to support ubiquitous connectivity via quality-of-service (QoS) provisioned applications and services. This aim requires that 5G should offer 1000x of 4G capacity, very low latency and high efficiency in terms of cost and energy [5]. Towards this direction, novel technologies that enable the mobile network operators (MNOs) to leverage the network infrastructure and the available spectrum resources have been already developed, such as the device-to-device (D2D) connectivity that allows the direct communication among closely located smart mobile devices or the network virtualization that enables the softwarization and slicing of network resources. Additionally, novel business models that enable the MNOs to combine their resources by establishing network sharing agreements have appeared. The technological evolution of mobile networks has also brought into the spotlight various cutting-edge OTT services, delivered by the OTT service providers (OSPs) that deploy

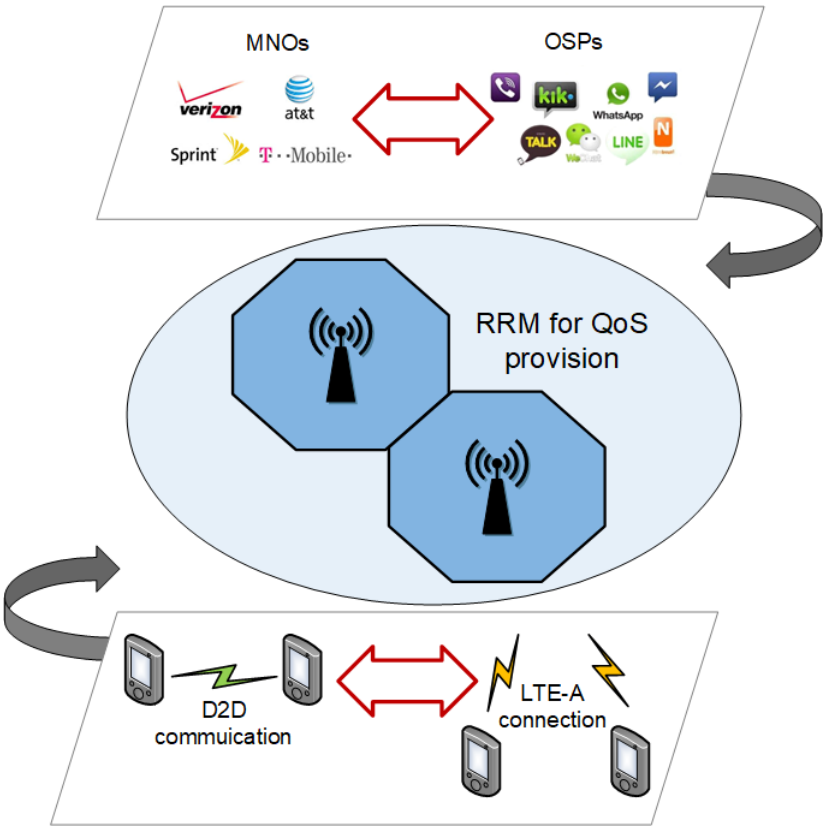


Figure 1.1: The 5G mobile network ecosystem

their own internal policies regarding QoS provision to their users. Although the various technologies and business paradigms may increase the network capacity and improve the QoS, from a technical aspect, advanced radio resource management (RRM) techniques for the efficient utilization of mobile network resources are required.

The picture that is slowly forming in view of the aforementioned trends is that of a multifaceted networking ecosystem where RRM is under the influence of the various wireless technologies with different capabilities and restrictions (e.g., LTE-A and D2D) and the various stakeholders (e.g., MNOs and OSPs) with different business policies (Fig. 1.1), whereas its outcome, i.e., the QoS experienced by the mobile users, is shaped by the interactions among different technologies and business stakeholders. Despite the fact that RRM has been extensively studied from the viewpoint of MNOs that aim to manage the LTE-A network resources efficiently, the challenges of modern RRM design stem from the interplay between LTE-A and other technologies, such as D2D, and between MNOs and other business stakeholders, such as OSPs, which may wish to intervene in RRM. The complexity of having to consider all of these interactions in the design of RRM techniques highlights the necessity of studying them separately in order to identify their characteristics and incorporate them to the RRM techniques accordingly.

Recently, the networking paradigm of D2D communication has gained considerable momentum. It leverages the proximity of UEs in a mobile network to directly route data traffic among them, alleviating part of the network load, as the data do need to traverse

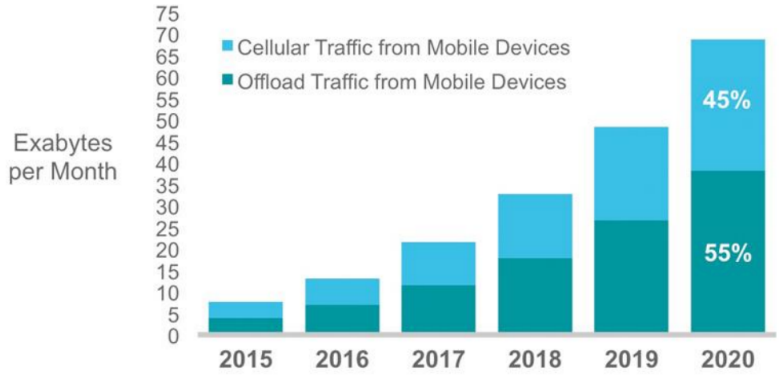


Figure 1.2: Mobile data traffic and offload traffic by 2020 [1]

the base stations and the core network [6]. The D2D concept has been investigated as a means for offloading the connections from the mobile network and improve the network performance in terms of spectrum efficiency and throughput. The D2D data exchange is inherent in mobile networks where the users communicate via mobile services that involve content sharing, gaming or social networking. The D2D connections can either operate in the licensed frequency band along with cellular communication links (inband D2D) or in the unlicensed spectrum, by utilizing wireless technology standards, like Wi-Fi, Wi–Fi Direct or Bluetooth (outband D2D).

Remarkably, the mobile traffic offloaded to Wi-Fi from cellular networks occupied 51% (3.9 exabytes/month) of total mobile traffic in 2015 and is expected to reach 55% (38.1 exabytes/month) by 2020, foretelling the prevalence of the Wi-Fi based outband D2D communication (Fig. 1.2). The outband D2D connections can be coordinated by the users and the neighboring UEs can be interconnected creating D2D communication pairs, share resources and relay received information through the Wi-Fi frequencies [7]. The UEs with both cellular and Wi-Fi interfaces can simultaneously maintain both types of connections. Moreover, the UEs may receive data via direct LTE-A connections and exchange the desired content fractions by establishing bidirectional flows. However, the D2D QoS aggravates when D2D links are of poor quality. In this case, the use of adjacent UEs as relays for the communication of pairs of UEs that exchange data can be a solution. The criteria for the selection of UEs as relays can be related not only to the channel status among the UEs but also to the users' social features, as socially connected users are more likely to engage in D2D cooperation [8].

As simultaneous transmissions over the same spectrum may lead to collisions that impact on the D2D QoS, RRM techniques for outband D2D communication are required for the sharing of spectrum among different UEs. More specifically, efficient medium access control (MAC) mechanisms that coordinate the UE pairs and their relays are required. The performance of D2D cooperative bidirectional communication can be improved if proper relays are selected. Additionally, the performance of D2D communication that involves the exchange of data concurrently received via the cellular links of the UEs may be affected by the mobile network characteristics, such as the scheduling policy and the

downlink channel conditions [9]. Thus, the outband D2D MAC performance should be studied considering the influence of the mobile network parameters.

Along with the technological advances, the mobile network ecosystem has also evolved in view of the introduction of innovative business models that enable the MNOs to accommodate high traffic volumes without the inflation of operational and infrastructure costs [10]. As the new deployment or upgrade of network infrastructure and the acquisition of additional spectrum is not always sustainable, the MNOs are able to actively share their existing network resources by establishing contractual agreements that regulate their cooperation. This business model has motivated the virtualization of mobile networks, as flexible network resource management techniques that implement the network sharing concept are required. In a virtualized network, the different tenants can obtain virtual slices (VSs), i.e., resources in virtual networks created dynamically on top of the same physical mobile network [11].

Apart from the MNOs that typically constitute the mobile network managers and owners, providers that offer OTT services, such as video streaming, teleconferencing or content sharing applications that complement the networking services offered by the MNOs, have entered the telecommunication industry [12]. Notably, in 2016, OTT traffic surpassed regular mobile traffic for the first time, highlighting the transition from voice and message services to OTT mobile services (Fig. 1.3). The OSPs offer services that have different QoS requirements, depending on their data traffic type, e.g., video delivery requires high bandwidth, while VoIP needs low latency, and may involve different user categories (e.g., free users or premium users paying for advanced QoS). However, the users access the OTT content through their mobile devices and their Internet access is often based on cellular connectivity provided by the MNOs. Therefore, the MNOs have full control of the network resources allocated to the UEs that generate OTT application flows and the OSPs are not able to intervene in order to accommodate their users' QoS demands in accordance with their user policies. The OTT application flows are considered to be of equal importance and are currently accommodated in a best-effort manner, which is not always acceptable by the OSPs, as they have their own performance goals, e.g., regarding the achieved grade-of-service (GoS) levels.

Despite that the intervention of OSPs in the RRM is a desirable feature that would enable them to apply flow prioritization policies, the resources may not be shared fairly among the OSPs, creating concerns about the network neutrality [13]. Although prioritization policies are necessary in certain cases, e.g., for gaming applications with low latency requirements, the resources should be accessed in an impartial manner, without monopolizing their utilization by some OSPs only. Therefore, the OSP-oriented RRM should balance flow prioritization and fair access to the MNOs' network resources.

Although the virtualization capabilities of modern mobile networks could enable the OSP intervention, the LTE-A standard does not specify a mechanism that would allow the dynamic interaction of the MNOs and the OSPs. As the VSs comprise of end-to-end network resources that refer to the radio access network (RAN) and the core network

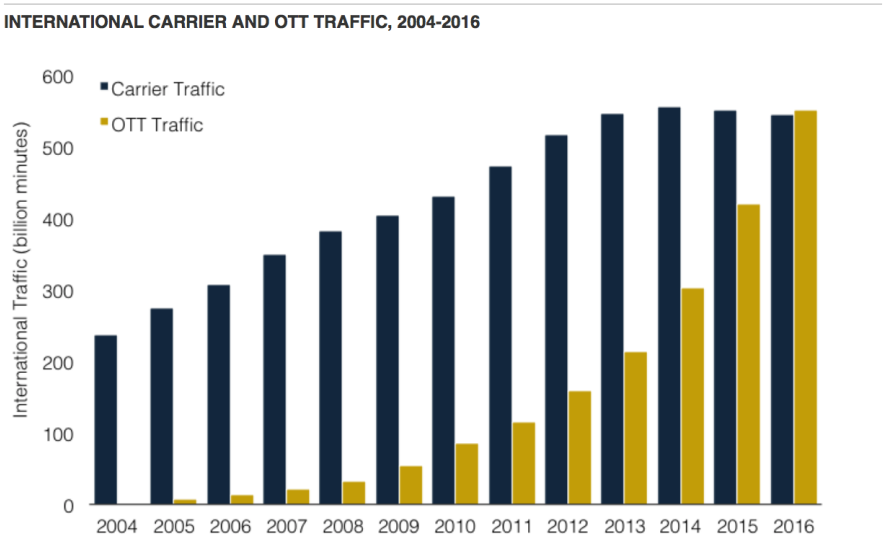


Figure 1.3: Mobile traffic vs OTT traffic [2]

(CN) and are allocated periodically, the OSPs should be able to periodically interact with the MNOs in order to obtain the required VSs in a timely manner. Hence, the RRM should be performed in a way that both OSPs' needs and the need for network neutrality are considered and within a reasonable time frame that will keep the delay experienced by the OTT application flows in acceptable levels.

Taking into consideration the aforementioned context, the 5G mobile networks are poised to be multi-tenant systems that rely on different networking technologies and involve various business models and stakeholders. Their interactions require efficient RRM techniques that conform to the multifaceted mobile network ecosystem. To that end, this thesis elaborates from a technological and a business aspect on the RRM design issue that arises in current LTE-A mobile networks and affects the experienced QoS. More specifically, two prominent networking technologies, LTE-A and D2D, are used for the study of technological interactions that affect the RRM design, whereas the interplay between the MNOs and the emerging OSPs constitutes the basis of the RRM study from the business viewpoint. Therefore, we first investigate the RRM issue for outband D2D communication from the MAC perspective, developing techniques that take into account the context of D2D cooperation and studying the influence of the mobile network on D2D connections. Next, we lay our focus on the RRM from the OSPs' perspective and develop a VS allocation method that enables the OSPs to request network resources in order to meet their users' QoS demands, applying flow prioritization policies without causing the violation of the concept of fairness among OSPs dictated by the network neutrality rules. The structure of the thesis and the contributions are discussed in detail in the following section.

1.2 Thesis layout and contribution

As explained in the previous section, the motivation of this thesis has stemmed from the RRM issue that should be addressed in the upcoming generation of mobile networks in order to meet the 5G goals for high QoS. Therefore, two research directions have been followed: i) the first direction refers to the outband D2D MAC design in the LTE-A context, and ii) the second direction focuses on the VS allocation for OSPs that aim to achieve QoS differentiation. The subsequent chapters provide details on the work that was elaborated for this thesis.

In Chapter 2 we provide the background related to the particular problems studied in this thesis. Particularly, the first two sections refer to the outband D2D communication in the LTE-A based mobile networks. The LTE-A network design and the D2D communication characteristics are presented. Subsequently, the issue of outband D2D MAC design is described in detail, along with the related work. The next two sections that comprise this chapter are related to the RRM issue for OSPs. The concepts that enable multi-tenancy in mobile networks are presented, along with the context of OTT applications. Next, the issue of resource management for OSPs is discussed, together with the existing works that are closely related.

Chapters 3 and 4 present our contribution related to the first research direction that focuses on the outband D2D MAC design. In Chapter 3, motivated by the MAC issues that arise in outband D2D communication, a new protocol that is based on the network coding (NC) technique, i.e., the adaptive cooperative network coding based MAC protocol (ACNC-MAC), is presented. Its performance is described by analytical models and evaluated under different D2D network scenarios. Moreover, the impact of LTE-A network deployment on the performance of the protocol is extensively studied via analysis and simulations. Our study has demonstrated that ACNC-MAC improves significantly the D2D throughput and energy efficiency comparing to NC-based SoA protocols, even in cases of high traffic load. More specifically, in scenarios of high downlink data rates, the induced throughput is up to 226% higher, whereas the energy efficiency is up to 38% higher. In Chapter 4, we discuss the integration of social awareness in D2D communication and explore various D2D cooperative networking scenarios. We next present the social-aware cooperative D2D MAC (SCD2D-MAC) protocol and evaluate its performance. We conclude our study by discussing the practical issues that may arise in the integration of social awareness in D2D cooperation. Our simulation results have shown that the integration of the social information of the users in the D2D MAC design improves the relay selection strategy. With the SCD2D-MAC protocol, up to 35% higher energy efficiency is achieved, whereas the average battery drain of the mobile devices that participate in the D2D cooperative communication is up to 58% lower, comparing to the SoA.

In Chapter 5, we focus on the second research direction and study the mobile network resource management issue from the OSPs' viewpoint. The outcome of this study is an OTT application flow prioritization algorithm that is based on the mathematical framework of matching theory and is described in detail in this chapter. Furthermore,

we provide analytical models that enable the thorough study of the performance of the proposed matching theoretic flow prioritization (MTFP) algorithm in various networking scenarios. We extensively evaluate the performance of the matching theoretic algorithm providing analytical and simulation results. Our study has shown that the OSP intervention in resource management with the aid of the matching theory can significantly improve the GoS, delay and energy efficiency levels, considering the priorities of the flows without violating the network neutrality rules. The MTFP algorithm achieves up to 50% lower GoS, up to 60% lower delay and up to 96% higher energy efficiency comparing to the best effort approach.

Finally, Chapter 6 concludes this thesis by summarizing the main contributions and sketching out possible research directions for future investigation.

1.3 Research contributions

In the process of elaborating this thesis, the following publications have been produced:

Journals

[J1] **E. Datsika**, A. Antonopoulos, D. Yuan, and C. Verikoukis, "Matching Theory for Over-the-top Service Provision in 5G Networks", *IEEE Transactions on Wireless Communications*, vol. 17, no. 8, pp. 5452-5464, August, 2018.

[J2] **E. Datsika**, A. Antonopoulos, N. Zorba and C. Verikoukis, "Software Defined Network Service Chaining for OTT Service Providers in 5G Networks," *IEEE Communications Magazine*, vol. 55, no. 11, pp. 124-131, November, 2017.

[J3] **E. Datsika**, A. Antonopoulos, N. Zorba, and C. Verikoukis, "Cross-Network Performance Analysis of Network Coding Aided Cooperative Outband D2D Communications", *IEEE Transactions on Wireless Communications*, vol. 16, no. 5, pp. 3176-3188, May, 2017.

[J4] **E. Datsika**, A. Antonopoulos, N. Zorba, and C. Verikoukis, "Green Cooperative Device–to–Device Communication: A Social Aware Perspective", *IEEE Access*, vol. 4, pp. 3697-3707, 2016.

International conferences

[C1] **E. Datsika**, A. Antonopoulos, N. Passas, G. Kormentzas, and C. Verikoukis, "Green Resource Management for Over-The-Top Services in 5G Networks using Matching Theory", accepted for publication in *IEEE International Conference on Communications (ICC)*, Kansas City, USA, 2018.

[C2] **E. Datsika**, A. Antonopoulos, N. Zorba and C. Verikoukis, "Matching Game Based Virtualization in Shared LTE-A Networks," *IEEE Global Communications Conference (GLOBECOM)*, Washington DC, USA, 2016.

[C3] A. Esfahani, G. Mantas, V. Monteiro, K. Ramantas, **E. Datsika** and J. Rodriguez, "Analysis of a Homomorphic MAC-based scheme against tag pollution in RLNC-enabled wireless networks," *IEEE International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, Guildford, UK, 2015, pp. 156-160.

[C4] **E. Datsika**, A. Antonopoulos, N. Zorba, and C. Verikoukis, "Adaptive Cooperative Network Coding Based MAC Protocol for Device-to-Device Communication", *IEEE International Conference on Communications (ICC)*, London, UK, 2015, pp. 6996-7001.

The contents that concern publications [J1], [J3], [J4], [C1] and [C4] comprise the body of this thesis.

Chapter 2

Background and arising resource management issues

2.1 Introduction

As explained in the previous chapter, this thesis focuses on the RRM design issue for mobile networks, elaborating on two different scopes:

(i) the scope of MAC design for outband D2D communication in LTE-A mobile networks

(ii) the scope of the OSP-oriented VS management in SDN-based shared LTE-A mobile networks

To that end, we will thereupon provide the technical background that may facilitate the understanding of this work and we will describe in detail the RRM issues under study.

The remainder of this chapter is structured as follows. First, in Section 2.2, we begin by providing an overview of the LTE-A network architecture, which is the basis of this work, and describe the process of radio resource scheduling in LTE-A based mobile networks. The Sections 2.3 and 2.4 are related to the part of the thesis that focuses on the outband D2D MAC design. Particularly, in Section 2.3, the concept of D2D communication is presented and in Section 2.4, the integration of social awareness in D2D communication is discussed. In these two sections, the D2D networking scenarios and the arising issues that have motivated our work on the D2D MAC design are analyzed. The subsequent sections, i.e., Sections 2.5 and 2.6, provide the necessary background related to the resource management from the OSPs' viewpoint. Section 2.5 describes the concepts

of network sharing, wireless network virtualization and network slicing that form the basis of multi-tenancy in mobile networks and presents the capabilities of the SDN framework. In Section 2.6, the characteristics of OTT applications and the OSPs are provided and the concept of network neutrality is explained. In the same section, the challenges of resource management for OSPs, which constitute the motivation of the development of the matching theoretic flow prioritization algorithm, are thoroughly discussed.

2.2 Mobile network design under the LTE-A standard

As wireless networking technology progresses, mobile networks have been significantly improved over the years in terms of QoS performance via a series of technological updates. The various telecommunications standards are developed and organized by the 3rd generation partnership project (3GPP), a mobile communications industry collaboration. In 2012, 3GPP has specified the long term evolution standard (Release 8) that builds upon the global system for mobile communications (GSM) and the universal mobile telecommunications system (UMTS) and improves the capacity of the third generation (3G) radio access networks, introducing novel digital signal processing methods and upgrading the network architecture to an Internet protocol (IP) based design with much lower latency [14]. The commercialization of the LTE technology had begun by the end of 2009.

Significant improvements on the LTE standard have been brought with the Release 10 that supports the use of a higher number of antennas and the feature of carrier aggregation with the aim of further increasing the achieved data rates and the stability of the LTE based networks, meeting the requiremens of the international telecommunications union (ITU) for the fourth generation (4G) telecommunication networks [15]. This upgraded version, noted as LTE-advanced (LTE-A), offers better cell spectral efficiency and 1 Gb/s downlink throughput. Many MNOs worldwide have adjusted their network deployments to the LTE-A specification, preparing for the upcoming fifth generation (5G) of wireless networks, which is currently under design by 3GPP.

2.2.1 LTE-A mobile network architecture

A contemporary LTE-A mobile network comprises of two main managing entities: i) the evolved UMTS terrestrial radio access network (E-UTRAN) and ii) the evolved packet core (EPC) network, as illustrated in Fig. 2.1.

The E-UTRAN is the radio access network (RAN) that offers the cellular connectivity to the UEs[1]. It refers to the deployment of base stations in an area, i.e, the evolved NodeB base stations (eNBs) for the macro-cells and the home eNBs (HeNBs) for the small cells, e.g., femto-cells and pico-cells. The eNBs are responsible for basic RAN functionalities, such as radio resource allocation, admission control and handover management, and can

[1]The terms mobile users and UEs are used interchangeably in this work.

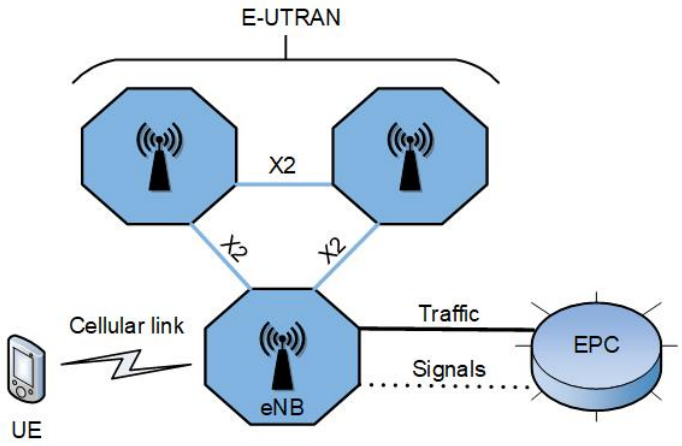


Figure 2.1: High-level LTE-A network architecture

communicate with each other via the X2 interface. They are also connected with the EPC via the S1 interface.

The EPC is the IP based core network (CN) that manages the E-UTRAN and consists of three types of components: i) the control entities that determine various control policies for the connected UEs (e.g., QoS and charging, resource allocation policies, etc.), ii) the gateways that forward the UEs' traffic and ensure that the packet routing, the UE admission and mobility control and the provided QoS conform to the policies imposed by the control entities, and iii) the subscription data entities that store the subscription profile of the connected UEs, providing the required information for UE authentication. The EPC is able to communicate with external IP networks, i.e., packet data networks in the Internet.

2.2.2 Radio resource scheduling for cellular connections

In the LTE-A RAN, the UEs connect to the eNBs that cover an area by establishing cellular connections, which are based on frequencies ranges within the ultra high frequency band. The frequency bands that form the available LTE-A spectrum are in the range 452-5925 MHz and may be different in each country [16]. The eNB-UE communication takes place with the use of frames with duration equal to 10 ms, each consisting of 10 subframes (Fig. 2.2). Each cellular link carries both uplink (UE to eNB) and downlink (eNB to UE) traffic, thus, subframes for both directions can be defined. For the downlink transmissions, the orthogonal frequency-division multiplexing (OFDM) scheme is employed, whereas for the uplink transmissions, a pre-coded version of OFDM called single carrier frequency division multiple access (SC-FDMA) is used, which reduces the high peak to average power ratio induced by the regular OFDM. A subframe lasts for 1 ms and can use a portion of the available bandwidth (up to 20 MHz, in case that carrier aggregation is not employed), which is divided into subcarriers of 15 KHz spacing. Subcarriers are organized into resource blocks (RBs) of 180 KHz each, thus, 12 subcarriers define an RB, which the minimum spectrum allocation unit. Hence, an RB lasts for 0.5 ms (7 OFDM symbols),

a time period defined as slot. Two slots comprise one transmission time interval (TTI), equal to 1 ms. The number of bits that can be transmitted per TTI is indicated by the transport block size (TBS).

As mentioned in Section 2.2.1, the eNBs allocate the radio resources to the connected UEs. This allocation relies on the estimation of the downlink channel conditions of a UE, which is performed with the aid of the channel quality indicator (CQI) reporting. The CQI is a 4-bit integer transmitted by the UE to the eNB that indicates the experienced downlink signal-to-noise ratio (SNR) and determines the modulation and coding scheme (MCS) that can be used in order to achieve the highest possible data rate, considering a target block error rate. The higher the SNR, the higher the order of the used MCS, thus, the resulting data rate is higher. The selection of MCS and the number of RBs assigned to a UE correspong to a specific TBS value [9], which determines the downlink data rate. The MCSs supported by the LTE-A standard are the QPSK, 16QAM and 64QAM, with modulation order equal to 2, 4 and 6, respectively.

The LTE standard does not specify how the radio resource scheduling should be performed. In reality, each MNO that deploys a network is able to select its own optimized scheduling algorithm. As no specific policy has been standardized, various concepts for the LTE-A downlink scheduling algorithm have been presented in the literature [17]. The resulting data rate of each UE varies according to the scheduling policy. There exist schedulers with different network performance targets, e.g., maximum throughput or proportional fair scheduler, which may be utilized in different cases, e.g., different resource allocation strategy according to the traffic levels and QoS demands of connected UEs.

The scheduling methods may be channel-unaware, i.e., they may assume that the downlink channel conditions do not vary in time, such as the round robin scheduler that assigns an equal amount of resources to all UEs in every TTI or the blind equal throughput algorithm that allocates resources to UES that have been served with lower average throughput in previous TTIs. Despite their simplicity, these approaches do not adapt in time-varying channel conditions, thus, they are not usually preferred in realistic LTE-A networks due to the low QoS performance. This deficiency can be tackled with the use of channel-aware scheduling methods, which can exploit the CQI reporting functionality of LTE-A networks, such as the maximum throughput scheduler that assigns all RBs to the UE that can achieve the maximum throughput in each TTI, aiming to maximize the overall network throughput, and the proportional fair scheduler that aims to find a tradeoff between fairness to all UEs and throughput maximization, ensuring that even UEs with poor channel conditions receive resources at some point.

For further QoS improvement, scheduling methods that provide QoS guarantees, e.g., in terms of achieved data rates or experienced delay, have been also developed, and are based in mathematical models, such as game theory. Although they seem more suitable for QoS provision in LTE-A networks, they may induce higher computational overheads that may hinder the timely adaptation of resource allocation to the changes of network conditions. To that end, the MNOs have quite a few options regarding their resource

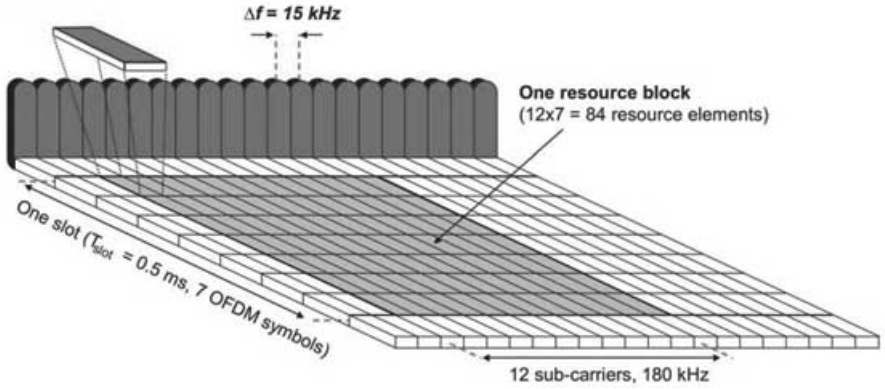


Figure 2.2: LTE-A time-frequency radio resources grid [3]

allocation strategies, while also being able to develop customized solutions, in order to meet their performance targets.

2.3 D2D communication in LTE-A networks

As the unprecedented growth of mobile data traffic has demonstrated the need for even higher network capacity, the MNOs seek novel approaches that will enable them to respond to the mobile users QoS demands. A promising idea seems to be the direct communication among UEs without the intervention of the RAN, known as D2D communication, which can be employed for offloading cellular traffic. It was first introduced in the LTE standard (Release 12) as a means of provision of proximity services (ProSe) related to commercial or public safety purposes and was considered as a part of the mobile network [18]. Furthermore, the D2D concept has been investigated in academia as a traffic offloading solution [19] and also by leader telecommunications companies, e.g., Nokia [20]. As new wireless standards emerge, like Wi-Fi Direct [21] and Millimeter-wave communication [22], which enable UEs' direct connectivity, integrating D2D communication into cellular networks seems appealing. In hybrid cellular/D2D networks (Fig. 2.3), the eNBs can exploit UEs' proximity and establish D2D links, increasing the spectral efficiency [23]. D2D connections operate in licensed frequencies along with cellular communication links, being under cellular control (*inband D2D*), or in unlicensed spectrum (*outband D2D*) [24].

In the case of inband D2D communication, the spectrum is shared among cellular and D2D connections, which is possible in two ways: i) the D2D links reuse spectrum portions used by cellular links (underlay D2D communication), and ii) the D2D links use the empty spectrum portions when they are not used by cellular links (overlay D2D communication). The inband D2D communication may significantly improve the experienced data rates, however, the interference induced by the sharing of spectrum resources among cellular and D2D UEs can be challenging and has been extensively studied in the last decade [25].

As an alternative to the use of cellular network resources, the outband D2D communication uses frequency bands of the unlicensed spectrum, such as the 2.4 GHz or 5 GHz

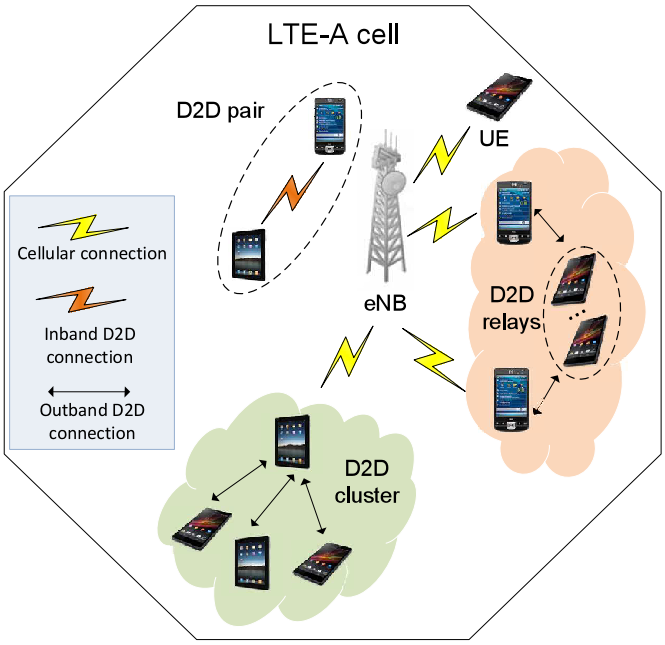


Figure 2.3: Cellular LTE-A network with D2D communication

band [26]. Only UEs with two wireless interfaces, i.e., LTE and Wi-Fi can support outband D2D connections, thus, the UEs can maintain concurrent D2D and cellular links. In this case, the connections among UEs can be either managed by the eNBs or can be totally autonomous, while they may be based on various wireless technologies. One very common option is the Wi-Fi technology that relies on the IEEE 802.11 standard. In this case, the UEs' connections are coordinated by the IEEE 802.11 medium access control (MAC) mechanism that uses the distributed coordination function (DCF), which is based on the carrier sense multiple access with collision avoidance (CSMA/CA) method [27], allowing the UEs to access the wireless medium in an ad hoc manner.

2.3.1 Cooperative outband D2D communication

The main advantages of the outband D2D communication are the unbinding of cellular system resources and the absence of interference from D2D connections to eNB-UE communication. Several works advocate for the use of D2D communication as a way to mitigate the limited network capacity problem. In realistic networks, the UEs' close proximity and the D2D data dissemination over Wi-Fi links create opportunities for UE cooperation. The formation of D2D networks can be an initiative either from the UEs or the cellular network. From the user's perspective, the coexistence of UEs that express their interest in downloading similar digital content from the eNB, e.g., video clips and advertisements, is typical in social activities, e.g., concerts or sports events [28].

Multiple neighboring UEs might desire to share multimedia content downloaded from eNB and create D2D clusters, e.g., using OTT applications related to content delivery and social networking. UE cooperation helps circumventing transmissions from eNB to

each UE separately, as devices can exchange data portions via Wi-Fi [29]. The UEs can share downloaded content fractions with peers via D2D bidirectional flows. Likewise, bidirectional D2D data dissemination can be performed by location-aware applications or multimedia services requiring information exchange between UE pairs, e.g., video telephony. The UEs' participation in collaborative clusters can be rewarded by the MNOs, making the cooperative transmission beneficial for both UEs requesting content and their helpers [30]. For instance, idle UEs with no interest in receiving specific content may be motivated to contribute as relays, if the operator provides incentives, e.g., lower service price or other types of remuneration [31].

Resource allocation can be performed using D2D clustering techniques and Wi-Fi Direct, combined with inter-cell interference control methods [32]. UE cluster formation can be a scheduling scheme, where only the cluster head receives data from the eNB and forwards them to the rest of the UEs [33]. However, the organization of clusters cannot easily adapt to volatile distributed topologies, and the assignment of the cluster head role to a particular UE, e.g., the one with the highest channel quality, is not fair especially regarding the energy consumption.

Besides being useful in user-oriented scenarios, D2D cooperation can be a solution to poor D2D link quality problems, which is a very common problem in wireless connections based on unlicensed spectrum. From the network's perspective, device collaboration can be facilitated by the exploitation of UEs as relays, i.e., cooperative UEs that are willing to retransmit data facilitating the communication of other UEs, when the UEs that attempt to communicate experience bad channel conditions.

2.3.2 Issues of outband D2D MAC design

Even though the inter-networking between cellular and D2D connectivity is apparent in a plethora of communication scenarios based on modern mobile applications, it is an aspect often neglected by the outband D2D schemes. As already described, in LTE-A networks, each eNB is responsible for the distribution of radio resources among the connected UEs, employing a variety of resource scheduling mechanisms. The utilized scheduling algorithm determines the achievable downlink data rates [17], which in turn regulate the packet arrival rates at UEs. As the resource allocation and transmission policies influence the frequency of packet arrivals, they further affect the QoS of D2D communication [9]. A joint methodology for user offloading to D2D network has been presented [34], which takes into account the interference among D2D links and captures the interaction between LTE-A and D2D connections. Despite its novel insights, this methodology does not consider the resource scheduling schemes and cellular data rates that cause differentiation in downlink performance among UEs.

In coexisting cellular and D2D networks, significant performance gains can be achieved by exploiting the devices' proximity, as UEs in the same area can act as relays and retransmit received and overheard packets. Conceptually, this store-and-forward process is related to the NC technique, which allows intermediate nodes to combine data from the

same or different information flows [35]. In D2D clusters formed by UEs that are connected through Wi-Fi links, the cooperation among devices can be leveraged by NC opportunities. The work published in [36] describes a scheme for data dissemination over D2D networks that exploits NC with the aim of improving the content availability at the UEs. This scheme regulates the data delivery considering the content correlation among neighboring UEs and utilizing the NC functionality for D2D data transmission. Nonetheless, it does not consider the dissimilar downlink data rates stemming from different cellular link states for each UE, as well as Wi-Fi related problems arising during D2D transmissions, two factors that result in unequal QoS provision at UEs.

Despite the improvement of LTE-A spectral use realized by offloading traffic to D2D links, the network congestion may be inherited to D2D communication level. Wireless channel access issues appear in the Wi-Fi based D2D clusters due to simultaneous channel contention from multiple D2D users (UEs or relay nodes). In unlicensed bands, Wi-Fi is the prevalent wireless technology adopted for D2D connectivity and is based on the IEEE 802.11 standard. However, with the densification of D2D networks and the increasing random access attempts by UEs, the utilization of IEEE 802.11 standardized MAC mechanism degrades the performance of cooperative transmissions. Furthermore, the time-varying quality of D2D links affects the throughput experienced by the UEs, as the packet losses, caused by bad channel conditions, increase the number of packet retransmissions.

Under the aforementioned circumstances, effective MAC mechanisms are required, which can improve the performance of the outband D2D communication. For several years now, NC has been widely utilized by MAC protocols, thanks to the throughput improvement it can achieve. This inherent capability can be further exploited by access schemes that manage D2D cooperative retransmissions [37].

So far, various NC-based MAC protocols have been presented in the literature. The seminal work presented in [38] provides the first practical implementation of NC for unicast traffic in wireless mesh networks, making use of two basic network capabilities: i) the opportunistic forwarding that allows each node to use only packets in its local queues for encoding, and ii) the opportunistic listening that enables each node to overhear packets communicated by its neighbors and use them for encoding decisions, exploiting the broadcast nature of the wireless medium. The proposed COPE protocol allows intermediate UEs along a path to apply the XOR operation to multiple packets, when the the intended next hop node has enough information to decode them. Considering the network of three nodes depicted in Fig. 2.4, COPE manages to improve the achieved throughput by 33%, reducing the number of transmissions from four to three for the transmission of packets a and b to their destinations 1 and 3, respectively. Nevertheless, it requires a history of received packets and their source nodes in order to form and send encoded packets to the nodes that can decode them.

Based on the main idea of the COPE protocol, the BEND protocol [39] combines packets at relay nodes, considering the union of all queue contents belonging to UEs

Figure 2.4: NC example using the COPE protocol

within a neighborhood. Any UE can encode and forward a packet even if it is not the intended receiver, in case that it senses that it can lead the packet to its destination. However, the creation of this "neighborhood coding repository" requires broadcasting of the UEs' queue status information.

Leveraging the existence of relays in the neighborhood of nodes that communicate in pairs, the NC based cooperative automatic repeat reQuest MAC (NCCARQ-MAC) protocol efficiently coordinates the cooperative transmissions [40]. Nevertheless, it allows cooperation only when NC conditions are met and assumes that the source nodes operate under saturated conditions, i.e., there always exist packets in their queues that are ready to be transmitted. This is not always the case in realistic D2D networking scenarios, where the packet arrival rates at UEs are determined by the LTE-A link performance.

Based on the IEEE 802.11 CSMA policy, the network coding aware cooperative MAC protocol (NCAC-MAC) employs a utility-based relay selection scheme that is able to find the relay with the bets channel conditions [41]. Using the NCAC-MAC protocol for D2D cooperative communication would require strict synchronization among UEs, along with a physical layer protocol that can handle information retrieval from corrupted packets, as it is assumed that the destination node has the capability to decode two packets a and b using only a corrupted version a' of packet a and the linear combination of a and b.

There also exist several physical layer NC (PNC) schemes that make use of the additive nature of simultaneously arriving electromagnetic waves. With PNC, the simultaneous transmissions by several source nodes result in the reception of a weighted sum of the signals by a receiver [42]. For instance, the work presented in [43] refers to a PNC based MAC protocol that targets bidirectional communication scenarios, whereas the work in [44] makes use of the multiple-input-multiple-output NC (MIMO-NC) principle that allows the relays to encode their own packets along with noisy versions of packets received from other nodes in the neighborhood. In spite of their traits, these schemes are not straightforwardly applicable to D2D networking, because they demand strict coordination of simultaneous transmissions.

Taking into consideration the characteristics of the D2D cooperative communication over the unlicensed spectrum, it can be observed that effective mechanisms are required for the resource management and the coordination of the channel access of the participating UEs. On one hand, the D2D MAC scheme should be able to leverage the opportunities for NC that may arise during the bidirectional communication of a pair of D2D users in both saturated and non-saturated D2D network conditions and also operate as a regular cooperative protocol when NC is not feasible. On the other hand, considering D2D communication scenarios where data are concurrently downloaded by the LTE-A network

(Section 2.3.1), the scheme is expected to operate under the influence of the LTE-A network parameters that determine the packet arrival rates at the D2D pair and influence the D2D MAC performance. To that end, in Chapter 3, we present the adaptive cooperative NC-based MAC protocol (ACNC-MAC) protocol that goes beyond the SoA outlined in Section 2.3.2 by prioritizing the transmissions of relays that are able to perform NC in saturated and non-saturated conditions. Moreover, we study its performance with the aid of a new analytical model that incorporates the LTE-A network parameters which affect the packet exchange rate at D2D level.

2.4 Integration of social awareness in D2D communication

Online social networking has offered unparalleled potential to communication among individuals in a plethora of everyday life activities. Indicatively, the 2014 football world cup in Brazil, apart from a sporting mega-event, was also a demonstration of social networks' power. In particular, 1.5 TB of data related to social media posts were circulated by the 75,000 spectators of the final match, corroborating the proliferation of mobile social networking nowadays [45].

Mobile users install a wide range of social networking applications on their devices and collaborate through them for personal and professional purposes. With the Wi-Fi capability of hand-held devices, users share their pictures with friends, edit and organize documents with co-workers or disseminate digital content to peers during social events. Thus, it becomes perceivable that D2D social networking involves communication among users that may or may not know each other, introducing different levels of trust among them. Moreover, it encompasses different application and content types and can lead to the formation of various cooperative networking topologies.

2.4.1 D2D social networking scenarios

From the networking perspective, cooperating users maintaining D2D connections create different network topologies that stem from communication flows among adjacent users. Naturally, physical proximity of users is a prerequisite for D2D networking. The D2D network structure varies depending on the location and the density of peers eligible for D2D communication. Apparently, a D2D network can be comprised of multiple pairs of users that share content fractions through bidirectional flows (D2D data exchange) or clusters, where users act as source nodes and transmit content to others (D2D content dissemination).

The differentiation of D2D use cases is better illustrated by contrasting the scenarios of a pair of colleagues jointly editing the same documents and a user group recording scenes of a football game and sending them to interested users in a stadium (Fig. 2.5). We thereupon describe D2D cooperative communication scenarios that occur when users interact using mobile applications, inducing different D2D network topologies.

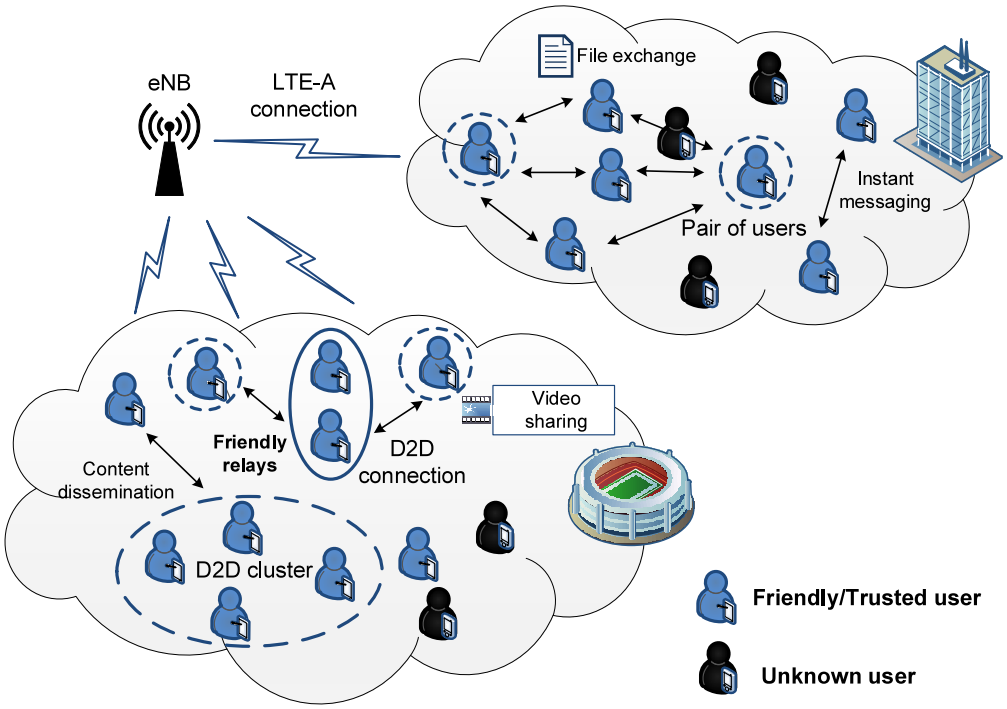


Figure 2.5: D2D social networking scenarios

2.4.1.1 Cooperative information exchange in D2D social networking

A common social networking scenario involves information exchange among users connected with interpersonal relations, such as friends on Facebook or colleagues on LinkedIn. People that already know each other are likely to share data when their devices are in Wi-Fi range.

D2D cooperation is realized at a personal level for content sharing between pairs of adjacent users, as shown in Fig. 2.5. Similarly, neighboring users exchange real-time information for specific purposes related to their location, e.g., a workplace [46]. In this case, multiple pairwise D2D connections coexist in the same premise or region. The exchanged information is private, as only the source and destination users are interested in it, whereas other users, might overhear the D2D transmissions. The existence of "friendly" users can be beneficial for D2D cooperation.

Lately, social D2D networking has expanded to mobile crowdsourcing applications, where cooperation is motivated by the existence of common goals, e.g., sharing live information on traffic conditions in order to or contributing to online communities. Another example is the technology of mobile augmented reality that enables mobile users to collaborate actively for the construction of accurate 3D models based on human perception of the environment. The D2D cooperative data exchange can also serve as an enabler of Machine-to-Machine (M2M) communication, which involves the interconnection of smart devices, usually without human intervention. The peer-to-peer model of D2D cooperation can support various smart applications, such as intelligent transportation systems or environmental monitoring [47].

2.4.1.2 Social cooperative D2D content dissemination

D2D cooperation can become an efficient means of content dissemination. Social events are a typical example of this scenario, given the high number of attendees and the coexistence of multiple devices in close proximity. Users are likely to be strangers but might belong to the same online community, usually related to their location, e.g., a stadium.

As depicted in Fig. 2.5, the attendees of a sports event share information with their peers through social media applications. Users can be organized in D2D groups, where some of them act as source nodes and the rest as destination nodes. The shared content is either user or cellular network originated. For instance, users transmit their own pictures to friends or other interested users or share videos previously downloaded via cellular connections. Users might also act as relays, supporting the content dissemination within groups of cooperating users.

As the density of users with social ties increases, more relay candidates that can assist the D2D transmissions of neighboring users exist. This feature could be useful in M2M communication scenarios, as a means to alleviate the cellular network congestion problem induced by the M2M links, using properly designed D2D cooperative schemes [48] or forming cooperative groups of "smart objects" that can improve the spectrum utilization [49].

2.4.2 Issues of social-aware D2D cooperative communication

D2D communication can mitigate cellular network congestion by exploiting users' physical proximity and offloading part of cellular traffic onto D2D links. Users' cooperation in this case is an initiative from the cellular network. D2D connections can be also initiated by users that wish to collaborate through mobile applications. Considering two devices within Wi-Fi range that interact through a social networking application, a D2D link can be established between them. Moreover, content sharing among community members in short physical distance can be realized by D2D cooperation.

The social D2D communication faces challenges similar to those of cooperative networking, which correspond to two basic questions:

(i) In which cases are D2D cooperative networks formed?

(ii) Which are the conditions for D2D cooperation?

Although these questions seem to revolve around networking issues only, they are accentuated by the social interactions of the mobile users.

The nature of mobile social networking stresses the need to consider the social factor when examining D2D cooperation. However, in realistic scenarios, the social awareness might induce energy consumption issues for the mobile devices that participate in cooperative transmissions. Hence, another crucial question arises: *how can social awareness be adapted to the green context of D2D cooperation?* In essence, the approaches that address issues of D2D cooperative networks should also have a green aspect that will enable them to utilize social awareness in a way that the D2D QoS is improved. Such an

D2D cooperation issue	Scheme	Mechanism	Use of social information	Energy consumption issues
Network formation	peer discovery [40]	classification of D2D users in groups and probing rate optimization per group	Users are separated in groups according to social centrality metrics.	UEs with high centrality in social networks are preferred for D2D cooperation leading to increase of battery usage. Uneven traffic distribution inside D2D coalitions may increase the energy consumption for highly preferred devices.
	cluster formation [41]	coalitional game theoretic distributed algorithm	Socially related users are organized in groups for cooperative video multicast.	
	D2D link establishment [42]	social-aware graph-based greedy algorithm	Users-members of the same community are interested in similar content but are reluctant to share content with users outside the community.	
Resource management	resource allocation in cellular and D2D users [44]	social graph based delay minimization algorithm	D2D users prefer to use the resources of cellular users in the same community.	Resource allocation according to social information only may result in unfairness in resource utilization and energy consumption among users.
	spectrum allocation in D2D users [45]	matching game with peer effects	D2D users with higher number of social ties are prioritized in spectrum allocation.	
Peer selection for relaying	probing scheme for relay selection [49]	optimal stopping problem formulation	Relay probing is performed according to users' social distances that define trust levels.	The D2D relay selection process based on social features may cause over-use of some devices as relays, unless proper energy aware strategies are employed or incentives for cooperation are provided.
	network-assisted relay selection [50]	coalitional game theoretic model	Users prefer relays they know (social trust) or agree to cooperate with strangers by assisting their transmissions (social reciprocity).	

Figure 2.6: Overview of existing social-aware approaches for D2D cooperation issues

extension would lead to higher D2D cooperative energy efficiency, which is an important performance metric when battery-driven devices are involved in D2D transmissions. A qualitative overview of existing approaches for the aforementioned challenges is shown in Fig 2.6.

2.4.2.1 Exploiting social features for green cooperative D2D network formation

At the D2D cooperative network formation phase, the peer discovery and communication mode selection issues arise, as users suitable to engage in D2D cooperation must be identified. In practice though, an adversity in D2D communication is the users' reluctance to cooperate by giving access to their devices to others or allowing the circulation of their own data via other devices [50]. These trust issues can be alleviated by offering the users proper incentives to "share" their devices with peers in area. Additionally to business level adjustments that motivate users to adopt D2D connectivity, social interactions among users can build a trustworthy D2D environment. Normally, users' mobile devices can maintain social preference lists that include contacts, namely friends, colleagues, etc., from installed social mobile applications. Thus, social characteristics can be easily extracted by social networking applications and serve as a guideline for D2D cooperative structures formation, e.g., using coalitional game theoretic models [8].

Users' social relations are long-term characteristics that can facilitate the discovery of proper D2D peers. The D2D candidate identification process can rely on information about the frequency of communication among users in communities [51]. Furthermore, centrality metrics characterizing the importance of users in social networks, can be utilized

for D2D network formation. In content sharing scenarios, as the one depicted in Fig. 2.5, several central users can be the source nodes and disseminate information to neighboring users-members of the same community. Instead of retrieving data from content providers, users can rely on data similarity in order to identify suitable peers, e.g., create user clusters for video multicast [52]. Under the assumption that users in the same community show interest in the same digital content, D2D connections can provide users in physical and social proximity with the desired content [53].

Similar social-aware rationale can be followed for users' communication mode selection. Users might not be eager to allow their devices to use D2D mode, even if D2D link quality is estimated to be higher than that of cellular link. Even though high peer density favors D2D networking, users' cooperation is finally endorsed in light of social factors, such as common desire for popular content and high trust among users, since the circulating data are of public interest (Fig. 2.5). Social information can be incorporated in cooperative game formulation, performing joint mode selection for sets of neighboring socially related users.

An aspect often neglected in existing social-aware approaches for D2D network formation issues is the energy consumption of participating devices. The D2D network design should incorporate energy-aware mechanisms that distribute equitably the traffic load among cooperating devices, without draining the resources of users with high centrality. For example, the MAC mechanism can allow cooperation only among peers with stronger social ties, e.g., peers that users contact more often. Moreover, energy efficient D2D cooperative structures could be established by enhancing D2D coalition formation approaches, such as [54], with social awareness.

2.4.2.2 Allocating resources for green D2D cooperation using the users' social information

Currently, LTE-A specification enables D2D connectivity over licensed or unlicensed spectrum. In the first case, cellular users coexist with D2D users. D2D links may share the same resources with cellular links or use dedicated spectrum. To mitigate the interference among the two types of links, resource allocation mechanisms can exploit social characteristics of D2D users. A point often overlooked is that the users are likely to show willingness for cooperation with their social connections but may not be interested in helping unknown users. Consequently, social-unaware spectrum allocation may become inefficient when users ignore opportunities for D2D cooperation.

In cooperative D2D networks underlaying cellular network, after deriving the social information, portions of spectrum can be allocated accordingly. For instance, users can be selected as cluster heads downloading content from the eNB and be allocated resource blocks in order to let neighboring users-friends receive data through D2D connections. In this way, spectrum efficiency is improved and eNBs fairly distribute available resources among users. Resource sharing can be performed among cellular and D2D users of the same community, in order to reduce the D2D transmission duration [55]. The resource

allocation problem can be also formulated as a matching game between users and spectrum resource blocks [56]. Bipartite graphs [57] could be also used for social aware resource allocation. As users with similar interests tend to request similar content, clusters are formed by users with strong social ties, increasing the number of requests offloaded to D2D links.

Particularly for D2D communication over unlicensed spectrum, D2D MAC protocols perform bandwidth allocation by coordinating channel access of multiple users. Examining a cooperative D2D clustering scenario, we see that if D2D transmissions of users with high centrality are favored by the MAC scheme, higher number of receivers is served at each communication round. However, the cooperation of a highly connected user as source node is hindered by the fact that it sacrifices his resources to benefit others. To that end, the exploitation of users' social interacting patterns could improve the performance of existing game theoretic MAC approaches [58].

Nevertheless, the existence of social ties is not a guarantee for users' willingness for cooperation. Given that the battery capacity of mobile devices is limited, there exists the contingency that spectrum allocation is not acceptable by the users due to high energy consumption. Therefore, D2D resource allocation schemes need to be energy-aware, considering at the same time the users' social information. Resource allocation and power control schemes that already exist in the literature, e.g., [59], could jointly consider the energy consumption factor with the social context and provide energy efficient D2D resource management.

2.4.2.3 Selecting socially connected users for energy efficient information relaying

D2D MAC schemes can promote D2D cooperation among socially related users by employing social-aware relay selection. Users desirable for relaying should gain channel access or be assigned spectrum resources with higher priority. As these preferences are defined in the social domain, the integration of social interactions in MAC design would reduce privacy concerns of D2D relaying, forming trustworthy cooperative D2D networks.

Once D2D pairs or groups are formed, the broadcast nature of the wireless channel enables opportunistic listening of circulating data fragments, creating fertile soil for D2D cooperation. Similarly to peer discovery, relay selection can be improved if users' social ties are exploited. A relay probing scheme can differentiate users with regard to both physical distance and social trust level [60]. In practice, users are more willing to assist the D2D communication of users that they know and trust than that of strangers [61].

The decision for cooperation is a dilemma between selecting suitable relays as mandated by wireless channel conditions and possible throughput gains, and preferring relays that maintain social ties with D2D users. Depending on the application, social trust may have higher priority than D2D performance, e.g., data exchange among colleagues in workplaces might have higher privacy demands than video sharing during sports events. This context information can be retrieved by eNBs and used for user-relay association in

matching theoretic tools with social-aware utility functions.

Arguably, when D2D trust is an issue, users with high number of social ties are mostly preferred as source nodes or relays. Nonetheless, even though cooperation may be beneficial for some users, it may result in relays' battery depletion. Proper relays that are able to assist the D2D communication of neighboring users should be selected, e.g., relays powered by renewable energy resources [62]. The energy consumption issue becomes more crucial in content distribution scenarios within D2D clusters, highlighting the need for incorporating the social information in energy-aware relay selection schemes. Furthermore, cooperation would be profitable for users eligible for relaying, if incentive mechanisms were applied by D2D cooperative schemes. Tangible profits, such as reduction of network service cost, could compensate for the energy consumed for relaying.

Considering the social networking scenarios discussed in Section 2.4.1 and the arising issues discussed in Section 2.4.2, it can be observed that the existence of social ties among users can be beneficial for D2D cooperative communication if the social awareness is introduced to the D2D MAC design. Towards this direction, in Chapter 4, we present a social-aware cooperative D2D MAC protocol (SCD2D-MAC) that is able to promote the cooperation among neighboring users that are socially related. It enables only the friendly users to act as relays that aid the D2D data exchange of a pair of users by creating a multicast group of friendly relays according to a list that indicates the social ties among the users provided by the eNB.

2.5 Multi-tenancy in mobile networks

Considering that modern mobile networks have become dense in sophisticated network elements and licensed spectrum is scarce and expensive, the provision of QoS to the mobile users dictates large-scale efficient resource coordination. As operating the infrastructure and obtaining additional spectrum is a significant expenditure for the MNOs, the need for reduction of the capital and operational cost has motivated the cooperation among MNOs, which allows them to gain access to new spectrum bands and network infrastructure, from base stations to backhaul network components, in a cost-efficient manner. Notably, it has been estimated that the cost savings can reach up to 30-40%, proving that pooling and sharing the infrastructure and licenced spectrum among different MNOs-tenants makes economic sense [63].

The MNOs are able to implement network sharing by cooperating for the management of different parts of the mobile network, as illustrated in Fig 2.7. A simple form of network sharing is the passive sharing that refers the joint management of eNBs, sites, masts and building premises by multiple MNOs. Aiming to further reduce the network expenditures, the MNOs have also attempted to share components that reside "deeper" in the mobile network, such as the active network components e.g., antennas, routers, switches and backhaul network equipment. The form of active sharing has further evolved and nowadays, it is feasible that the MNOs share also their spectrum, co-managing both

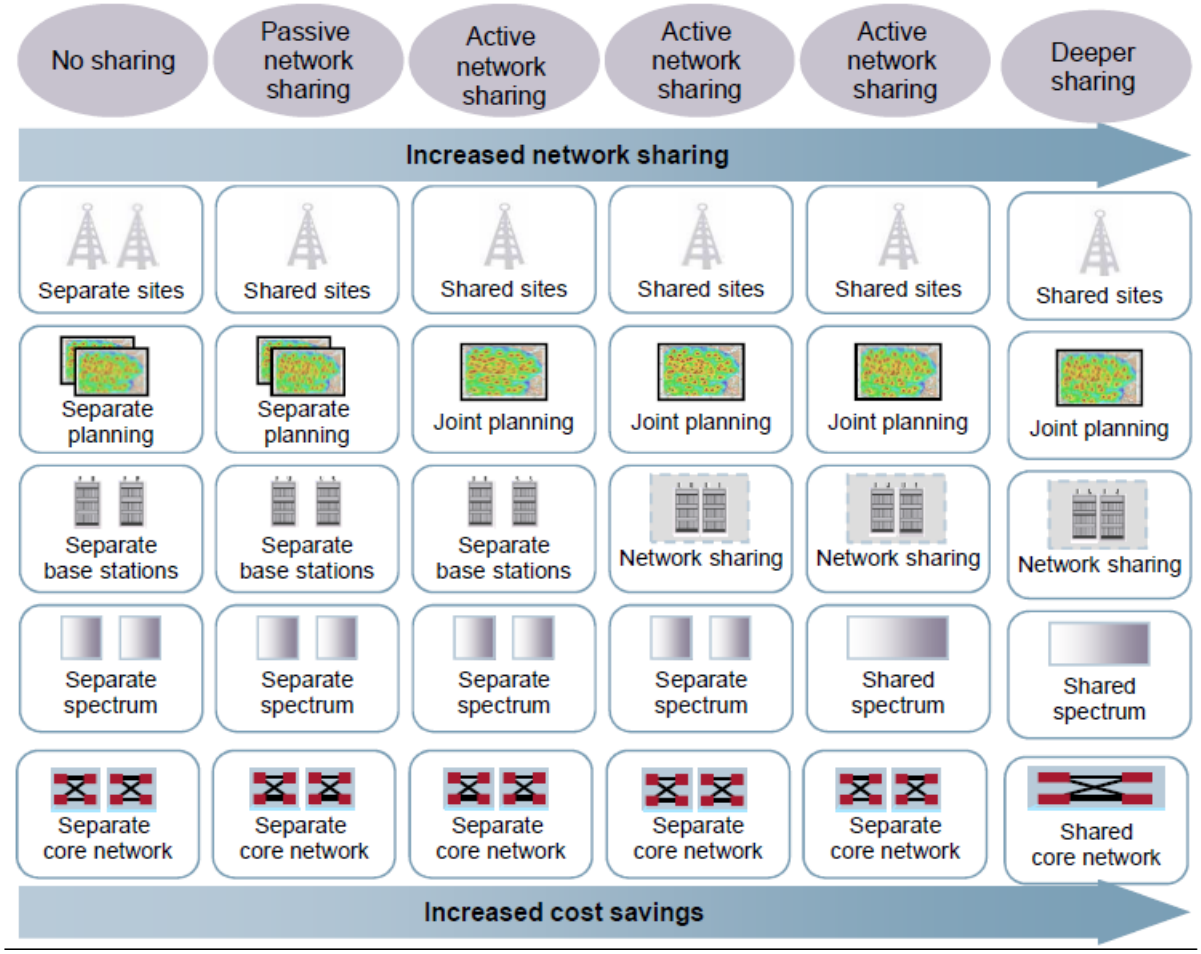


Figure 2.7: Types of network sharing (3G4G Small Cells Blog-smallcells.3g4g.co.uk)

the RAN and the CN.

2.5.1 Network sharing in LTE-A networks

The LTE-A technology can incorporate the novel business model of network sharing [64]. More specifically, two main network architecture designs that allow the management of an LTE-A based network, i.e., the RAN and/or CN elements, jointly by multiple MNOs, have been specified by 3GPP (depicted in Fig 2.8): i) the multi-operator CN (MOCN), where CN elements owned by different MNOs are connected to a shared RAN, and the gateway CN (GWCN), where the MNOs share the CN in addition to the RAN elements [65].

However, both MOCN and GWCN sharing configurations are conceptual and do not clarify the mechanisms and technologies that would be the basis of the operation of the shared network. The implementation of network sharing according to these architectures is left to be decided by the MNOs. To that end, the virtualization of resources (of both CN and RAN) is useful and can facilitate the access of all tenants to the complete set of available resources (network exposure), enabling them to create a shared pool of resources while also managing their part of the resources in isolation.

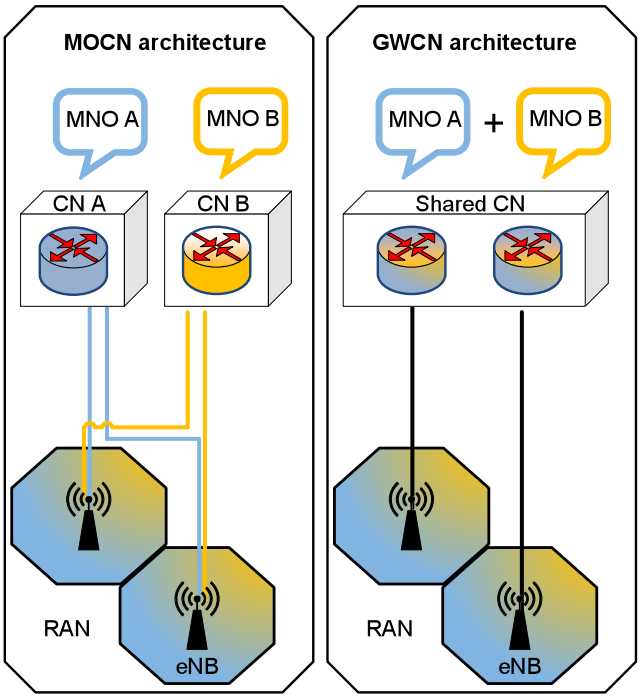


Figure 2.8: Network sharing architectures specified by 3GPP in TS 23.251

2.5.2 Wireless network virtualization and network slicing

Aiming to meet the users' QoS demands avoiding the inflation of the expenditures for the network management, the MNOs are able to share their infrastructure and spectrum, with the aid of network virtualization that abstracts and slices the network resources into VSs. The VSs are virtual networks that comprise of softwarized network functions, managed by different tenants in isolation [66]. Each VS includes end-to-end network resources in the RAN and the CN. The network slicing concept implies that actions performed in one VS do not affect the other VSs, even if they share the same underlying physical hardware. It is expected to be a key technology in 5G networks, as it facilitates the fine-grained control of network services and the flexible customization of the resources allocated to the different network tenants and their users thanks to the network exposure capability it offers.

The virtualization facilitates the network exposure with the aid of software defined networking (SDN) (Fig. 2.9) that provides controllers for centralized management of a programmable network and enables the network disaggregation, ensuring the isolation of VSs that may vary in time and belong to different tenants [67]. The SDN has been widely used in VS management methods, such as SoftRAN [68], Orion [69], OpenRAN [70], SoftAir [71], etc.

The SDN framework has emerged from the Openflow interface, specified by the open networking foundation (ONF), which enables the communication between the network switches and routers, i.e., packet processing machines and allows their management by a centralized controller [72]. Similarly, in SDN, a centralized software-based controller

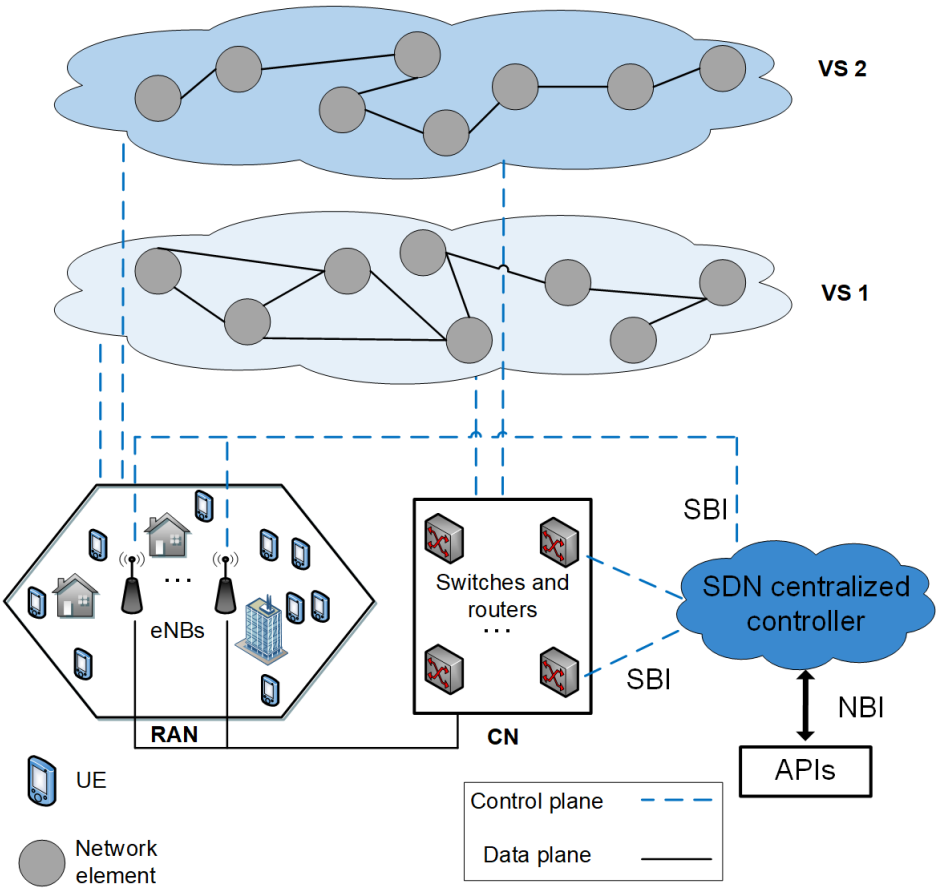


Figure 2.9: Network slicing with the aid of SDN framework

abstracts the network control plane and the underlying data forwarding plane, including both physical and virtual devices of the CN and RAN [73]. SDN also provides two types of application programming interfaces (APIs) that facilitate the network orchestration, i.e., the northbound interface (NBI) that enables the communication between the controller and the various network services and the APIs of the network managers (e.g., MNOs) and the southbound (SBI) interface that is used for the communication between the controller and the packet forwarding elements (switches and routers) of the network.

2.6 OSPs and OTT applications

The advances in wireless networking and smart mobile devices since the early 2000s have introduced a plethora of multimedia based mobile applications, such as YouTube, Netflix, Skype, Facebook, WhatsApp, etc., known as OTT applications. Indicative of how mainstream these applications have become is the observation that, in 2016, YouTube has reached 113 million users in North America [74].

The functionality of OTT applications is based on the Internet connectivity that is provided to the mobile users by the cellular networks MNOs, such as Orange, Telefonica, etc., or via Wi-Fi connectivity. Thus, the characterization over-the-top stems form the

fact that OSPs act as third party providers that deliver their services on top of the Internet, bypassing the regular Internet providers of mobile users, namely the MNOs that own and manage spectrum and network infrastructure. The OTT services mostly refer to communication services, such as voice, messaging and social networking, etc., and content delivery, such as video streaming. Each service may have different QoS requirements, e.g., low latency in teleconferencing applications or high data rates in video on demand. Moreover, different user categories may exist, such as premium users that pay for advanced usage privileges, freemium users that are charged only for additional proprietary features or free users. Each of these categories may correspond to different QoS performance levels even for the same OTT application, e.g., higher video quality for preemium users in YouTube[2].

From a more technical point of view, in modern wireless networks, the usage of OTT applications imply content downloading/uploading and content exchange among the mobile users. These content delivery operations generate OTT application flows carried through the wireless network infrastructure. The users are able to upload or download content via their cellular connection or exchange data directly among their mobile devices via Wi-Fi connection.

2.6.1 Interaction of OSPs and MNOs

Although the OTT application flows circulate over the MNOs' networks, there exist no business agreements between OSPs and MNOs. Hence, the MNOs are not in control or responsible for the distribution of the OTT content and merely deliver the content, without producing any revenue from acting as the Internet connectivity providers. Actually, the proliferation of OTT services has caused the loss of 386 billion dollars in revenue of MNOs from 2012 to 2018[3], threatening the role of MNOs as principal stakeholders in the telecommunications market. Furthermore, OTT applications that are similar with existing services of MNOs, e.g., WhatsApp and an MNO's text messaging service, have been introduced, disrupting the monopoly of MNOs.

Aiming to maintain their position in the market, the MNOs may develop different strategies. For instance, an MNO may attempt to obtain an already established OSP, develop its own competitive OTT services, restrict data usage for OTT services or create partnerships with existing OSPs under contractual agreements. The selection of each strategy is a result of a thorough cost-revenue analysis, however, the option for cooperation seems to be a beneficial solution when the investment risk is too high or the competition is too hard [75]. It has been reported that 327 partnerships have been established in 2015, as shown in Fig. 2.10[4].

The cooperation of MNOs and OSPs requires the regulation of the practical interaction

[2] https://www.cnet.com/how-to/youtube-red-details/

[3] http://fortune.com/2014/06/23/telecom-companies-count-386-billion-in-lost-revenue-to-skype-whatsapp-others/

[4] https://www.detecon.com/en/Publications/when-competitors-turn-partners

Figure 2.10: Yearly OTT partnerships (2010-2015)

among the two parties. On one hand, the MNOs that maintain control of their traffic need the monetization of OTT traffic that circulates using their mobile network infrastructure and spectrum in order to be able to produce revenues. On the other hand, the OSPs need a certain level of OTT application traffic management in order to be able to achieve the QoS requirements of the mobile users and provide them with properly customized services. The common goal of OSPs and MNOs is the provision of high quality services to the users, hence, their collaboration should be regulated in a way that the required performance levels are achieved for both parties.

The establishment of partnership deals implies the MNO-OSP interaction in a practical level, i.e., through proper cooperative managing of OTT application flows. The intervention of OSPs should be supported by the underlying network infrastructure, offering the necessary tools for the development and implementation of OTT user policies. Additionally, the MNOs should keep their role as network supervisors that ensure the application of the rules defined by the business agreements regarding the management of OTT application flows.

The MNO-OSP partnership bears some resemblance to the case of cooperation of MNOs with mobile virtual network operators (MVNOs), which do not own spectrum or infrastructure. The MVNOs are operators that seek to improve their services by leasing part of existing networks owned by MNOs, without having to deploy their own infrastructure. In this sense, OSPs have a similar goal, as they aim to increase the popularity of their applications, making high quality services available to more users. Nevertheless, the MVNOs are currently able to negotiate network resources with the MNOs, aiming to increase their capacity (for voice and data) and offer innovative bundles

and networking packages to the end users [76]. This means that the MVNOs are providers of cellular connectivity, similarly as the MNOs. In contrast, the OSPs are providers of totally different types of services, e.g., content delivery, social networking, etc., with more complex and time-varying QoS demands, related to different user categories. Moreover, an OSP may need to cooperate with different MNOs that cover an area in order to obtain the required network resources in short time scales that match the OTT traffic dynamics. Therefore, fine-scale and dynamic negotiation mechanisms are required for efficient cooperation among OSPs and MNOs. The development of such mechanisms remains an open issue for both academia and industry, in view of the culmination of the OTT application usage.

2.6.2 The network neutrality principle

Currently, the OTT services have no QoS guarantees and are provided in a best-effort manner. Aiming to offer high quality services, the OSPs may opt to collaborate with MNOs in order to be able to apply their user policies, obtaining the required network resources. This collaboration may induce preferential treatment of certain OTT users/OTT application flows, when the OSPs need to apply flow prioritization policies for certain users. As OTT services should operate under the network neutrality rules of the public Internet, the MNOs should not differentiate the OTT flows that circulate in the mobile networks [77].

The network neutrality concept has been introduced a few years ago as a regulatory framework for the interaction of content providers (CPs) and internet service providers (ISPs) [78]. Its main principle imposes that ISPs should not charge additional fees for giving priority to or improving the QoS of traffic from CPs that are willing to pay higher fees for better content delivery. Offering prioritized network services in return for fees may increase the cost of CPs' services, hindering the free access of the end users.

The network neutrality is very closely related to the case of MNO-OSP cooperation. Although the acquirement of network resources may improve the offered services, which is profitable for both OSPs and MNOs, it may also lead to undesirable prioritization of certain OSPs at the expense of others [79]. For the maintenance of an adequate fairness level among OSPs, it is required that no OSP's users and OTT application flows are prioritized in way that the access to network resources of other OSPs becomes restricted.

To that end, prioritization should be applied at OTT application flows and not at the OSP level, e.g., by prioritizing preemium users of different OTT applications in return of monetary reward, respecting the network neutrality rules. The MNO-OSP negotiation mechanisms should be able to maintain this balance, incorporating suitable policies that ensure fairness among the participating OSPs.

2.6.3 Issues of resource management for OSPs

The OTT application flows may have different QoS demands depending on the type of their data traffic, e.g., low latency for gaming applications or high data rates for video streaming. Moreover, each application may involve different user categories, e.g., freemium users or premium users paying for advanced usage privileges [80]. Hence, the flows are of dissimilar importance, determined by the corresponding OSPs' policies. In LTE-A networks, when VSs are created, the flows receive resources in a best effort manner, regardless of their priorities [81]. The OSPs are not involved in the VS allocation, thus, they do not control the QoS levels in terms of various performance indicators, e.g., GoS, and cannot apply flow prioritization when required, as MNOs fully control the UEs' connections and decide about the allocation of CN and RAN resources to the flows. To that end, enabling the intervention of OSPs in resource management might be profitable for both OSPs and MNOs [82], as delivering high quality services is a primary goal for both parties. The cooperation of OSPs and MNOs for the joint deployment of network infrastructure has demonstrated their common interests [83].

The OSPs' intervention in VS allocation requires that the network architecture enables the OSP-MNO interaction exposing the network services, e.g., via network APIs [84] and the use of SDN. Despite the availability of VS management tools, it is not clear how the resources are shared among OSPs with flows of different priorities. The resources should be shared impartially among applications, thus, prioritization should be applied at the flow level, while fairness should be guaranteed at the OSP level, as dictated by the network neutrality rules [85].

The VSs encompass resources of both the CN and the RAN, thus, end-to-end resources are allocated to OSPs' flows. The RAN resource allocation, i.e., association of flows with eNBs and spectrum allocation, is of fundamental importance for the flows' QoS [68], whereas the CN resources, i.e., bandwidth in CN links, should not be neglected. Specifically, RAN resource scheduling periodically allocates spectrum resources in UEs' cellular links. The spectrum allocation is adjusted in each VS allocation round according to network-related parameters (e.g., congestion of links, UEs' channel conditions, etc.), and MNOs' performance goals (e.g., spectral efficiency maximization, etc.). Given the periodicity of the VS allocation and the dynamic number of flows concurrently requesting resources, flows may not receive resources in each round, experiencing time delays during their service time. Moreover, when OSPs' policies are considered, the network coordinator (e.g., a centralized controller) should periodically interact with the OSPs. As information about the flows needs to be exchanged between the RAN and the OSPs, the CN links also experience congestion. Hence, the CN influences the delay of VS allocation not only regarding the time needed for the reception of required resources by the flows, affected by the RAN resource scheduling technique, but also regarding the time required for the transmission of flows' information through the CN. Despite that the existing resource allocation approaches could be applied as scheduling techniques in each VS allocation round, no insights for the delay they may induce have been provided.

Although the network slicing concept implies the allocation of resources both in the CN and the RAN [86], the vast majority of resource allocation approaches refer to RAN resources. Resource scheduling is performed either in a single evolved NodeB base station (eNB) (e.g., [17], [87]), or in a shared RAN, allowing the sharing of eNBs and/or spectrum resources among MNOs (e.g., [88, 89, 90]) or virtual MNOs (MVNOs) that do not own spectrum or infrastructure and lease VSs from the MNOs (e.g., [91, 92, 93, 94, 95]). Even though some of these schemes, mainly based on game theory, could potentially apply to OSPs, two main issues arise: on one hand, the OSPs need to prioritize certain flows according to their policies, whereas, on the other hand, the network neutrality concept opposes to the discrimination of OTT application content of certain OSPs. Hence, prioritization should be applied at flow level and also be impartial towards the involved OSPs. However, it is not clear how this type of prioritization can be incorporated in the existing schemes.

Moreover, ensuring that a regular optimization scheme adheres to the network neutrality property is not straightforward, as the integration of prioritization may arise fairness issues among the different OSPs. On the other hand, in order to derive tractable optimization problems with commonly utilized approaches based on game theory and most optimization schemes, the utility functions that describe the OSPs' performance goals should have specific structure. This condition does not always hold in performance metrics employed in wireless resource allocation methods, e.g., the GoS metric [96]. Additionally, an OSP would have to be aware of the other OSPs' policies in order to decide about its preferences, an information that is required by game-theoretic approaches for wireless resource allocation.

The VS allocation problem under study can be also considered as an asymmetric assignment problem where a number of objects (flows) have to be assigned to a smaller number of persons (eNBs), which can be solved by auction algorithms. However, these algorithms (e.g., [97]) require that the profits of the bidders (flows) and the prices of the sellers (eNBs) have specific properties that assure convergence to a feasible solution. It is not easy to incorporate the complex preferences of the different players in the strictly defined prices and profits. The considered problem also resembles the well-known problem of matching in bipartite graphs, i.e, the construction of a set of edges without common vertices. Given that the flows are characterized by different priotities and QoS demands, the VS allocation problem that we study resembles a variant of the graph matching problem that considers edges with different weights. Nevertheless, in our work, each OSP seeks to minimize the GoS, whereas the MNOs aim to minimize the average GoS of all OSPs. Thus, the weights of the edges would obtain different values depending on whether the preferences of the OSPs or the MNOs are considered. The difference in the preferences of the two sets of players cannot be depicted in the weights of the edges. It is difficult to define a single utility function that would describe the interaction between OSPs and MNOs and provide numerical values for the weights that would enable the resolution of this assignment problem with the aid of regular graph matching algorithms [98].

Given that a matching between flows and eNBs is the desired outcome of the VS allocation process, the mathematical framework of matching theory can be employed for the design of a resource allocation solution. One of the most well-known problems addressed by matching theory is the stable marriage problem, where a set of men and a set of women aim to select the most desirable spouse (one-to-one matching). In the model presented in [99], the preferences of the players of each set are represented by an ordered list of items that belong to the other set, e.g., a man's preference list consists of an ordered list of the most preferred women. However, the problem considered in our work is different from the stable marriage problem for two main reasons: on one hand, the association between flows and eNBs is a many-to-one matching, as several flows may be assigned to one eNB; on the other hand, the preferences of the players should incorporate the priorities of the flows and their QoS demands, enabling the OSPs to indicate their preferences over the network resources according to their user prioritization policies and the MNOs to allocate the resources to the flows in an OSP neutral manner, abiding by the network neutrality principle.

Taking into account the context of the considered VS allocation problem, it should be noted that it has common characteristics with the hospital-doctor matching problem described in [100], where a number of doctors seek to be matched with hospitals with the aim of achieving the highest possible wage or better working conditions, e.g., flexible working hours. This problem is modeled as a matching game with contracts that can express the preferences of the players of each set over the players of the other set and define the conditions, e.g., wage or working hours, that may characterize their association. In a similar manner, the matching framework could be adapted to address the considered problem in a way that the use of contracts enables the flow prioritization in the resource allocation, without violating the network neutrality rules.

An overall inspection of the related approaches for resource allocation problems in wireless networks shows that, despite their benefits, they are not applicable or may not be easily adapted to the considered VS allocation problem. They do not explicitly consider the co-existence of flows that belong to different OTT applications, thus, they do not provide a means for the OSPs to apply their user policies in a way that the network neutrality is respected. Also, they do not consider the delay experienced by the flows that experience a sequence of VS allocation rounds, which is affected by the CN congestion levels.

Chapter 3

The ACNC-MAC protocol for D2D communication

3.1 Introduction

As already described, in LTE–A networks, the UEs can concurrently participate in cooperative outband D2D data exchange by virtue of user- or network-related parameters (e.g., interest in the same content and cooperative transmissions, respectively). In these scenarios, two major problems arise: i) the coexistence of multiple devices creates channel access issues, demanding effective MAC schemes, and ii) cellular network factors (i.e., scheduling policy and channel conditions) affect the D2D communication, as the circulating information in D2D links is mainly of cellular network origination, requiring the study of cross–network interactions.

In joint cellular/D2D networks of the current and the upcoming generation, the communication among UEs induces the use of different wireless technologies. There exist the contingency that the UEs receive data from the eNB and concurrently share them through D2D links. The coexistence of different connection types entails cross-network interactions. In particular, the characteristics of the cellular network may affect the performance of the D2D connections. To that end, the D2D MAC scheme should be properly designed in order to leverage the NC opportunities that arise in the bidirectional communication of D2D pairs and its performance should be studied within the context of coexisting cellular and D2D connectivity.

Taking into account the aforementioned characteristics of the outband D2D communication in LTE-A based cellular networks, in this chapter, we present an adaptive cooperative NC-based MAC protocol (ACNC-MAC) protocol. More specifically, the contributions of this chapter can be summarized in the following points:

(i) *Design of the ACNC-MAC protocol:* ACNC-MAC allows neighboring UEs to act as relays and perform cooperative transmissions, assisting a UE pair's D2D communication. It goes beyond the SoA protocols by better exploiting NC opportunities arising in bidirectional D2D communication. The relays that overhear packets from both UEs transmit encoded packets, serving both flows at each communication round. The proposed protocol prioritizes the transmissions of relays that are able to perform NC, maximizing the benefits of cooperative D2D communication.

(ii) *Throughput analysis of the ACNC-MAC protocol in saturated conditions:* We provide an analytical model for the achieved D2D network throughput in saturated conditions, when the ACNC-MAC protocol is employed. The proposed analysis effectively incorporates the ACNC-MAC rules and models the ACNC-MAC throughput performance considering the number of available relays and the packet error probabilities in the D2D links.

(iii) *Cross-network analysis of throughput performance of ACNC-MAC:* As the UEs that engage in D2D communication simultaneously receive the desired content from the eNB and share it with their peers, we study the D2D MAC performance in the LTE–A context. Particularly, the packet exchange rate at D2D level is dictated by the packet arrival rates at the UEs, which are affected by i) the downlink resource scheduling policy, and ii) the UEs' cellular downlink channel conditions. Considering these cross–network interactions between LTE–A and D2D communication levels, we propose the incorporation of cellular link parameters in the analysis of D2D MAC performance. Specifically, we present and validate an analytical model for the D2D throughput achieved by ACNC–MAC that captures the LTE–A parameters.

(iv) *Evaluation of ACNC-MAC performance under the influence of concurrent cellular and D2D connectivity, simulating realistic scenarios:* We perform extensive simulation study in both saturated conditions and non-saturated conditions with varied traffic rates. Moreover, we study the impact of specific LTE-A network characteristics, i.e., the downlink transmission scheduling policies, the downlink channel conditions and the cell congestion levels on D2D MAC performance and the ACNC-MAC behavior in the LTE-A context. Recognizing the escalating demand of digital video by mobile multimedia applications, we also evaluate the proposed protocol in D2D video transmission scenarios, where UEs exchange video content downloaded through cellular connections.

The remainder of the chapter is structured as follows. In Section 3.2, the considered system model is described and in Section 3.3, a detailed description of the ACNC-MAC

protocol is provided. In Section 3.4, an analytical model for the achieved D2D network throughput using ACNC-MAC in saturated conditions and an analytical model for the D2D throughput of ACNC–MAC that captures the LTE–A parameters are presented. In Section 3.5, the proposed analysis is validated and the performance of the ACNC-MAC is evaluated in different simulation scenarios. Finally, the chapter is concluded in Section 3.6.

3.2 System model

We consider a single-cell cellular network with one eNB and K UEs in the area of cell coverage (Fig. 5.1). Each UE is equipped with two radio interfaces, LTE-A and Wi-Fi, thus they are able to maintain connection to the eNB and simultaneously connect to other UEs via Wi-Fi. The cellular and D2D transmissions use different frequency bands. More specifically, the cellular transmissions utilize the frequency bands that form the available LTE-A spectrum in the range 452-5925 MHz, whereas the D2D transmissions use unlicensed frequency bands, e.g., the IEEE 802.11 2.4 GHz band [101]. The UEs UE_1 and UE_2, denoted as *active UE pair*, request content from the eNB and establish LTE-A connections. Packets p and q arrive at UE_1 and UE_2, respectively, through cellular connections. The two UEs are interested in each other's received content and they wish to establish bidirectional links among them.

As the considered D2D network coexists with the cellular network, the D2D communication experiences challenges related to the interaction between LTE-A and Wi-Fi. When D2D communication involves data concurrently downloaded by the mobile network, QoS of D2D connections may be affected by LTE-A network parameters, such as the resource scheduling policy and the downlink channel conditions. These parameters determine the data rates achieved in the downlink channel and the induced packet arrival rates may affect the D2D performance. Therefore, we next describe the characteristics of both types of connections jointly considered in our work.

3.2.1 LTE-A communication (eNB to UE)

The design of the LTE-A downlink physical layer is based on the OFDM scheme that allocates specific patterns of subcarriers in the time-frequency space to different users, as described in Section 2.2.2. The OFDM symbols required for the downlink transmissions are organized in N_{RB} RBs. A portion b of N_{RB} in the time-frequency domain is allocated to each UE by the eNB in every TTI.

3.2.1.1 Downlink resource allocation

The requirements of each UE regarding the allocated RBs stem from PHY layer parameters of the eNB-UE connection that reflect the LTE-A link quality, i.e., the SNR levels and the employed MCSs. In our work, without loss of generality, we use a round robin sched-

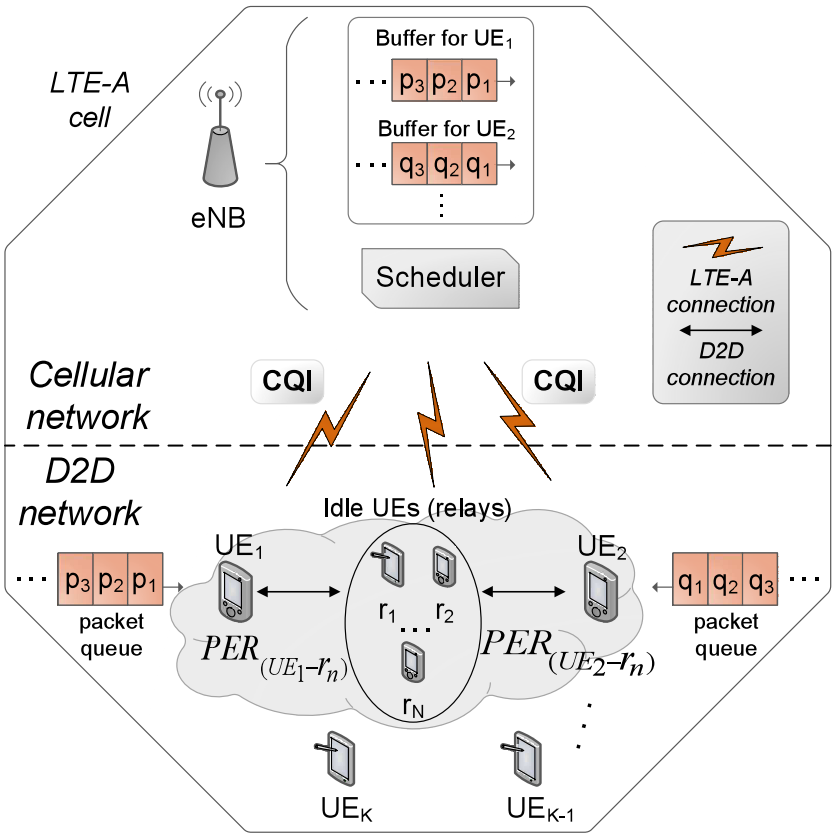


Figure 3.1: D2D enabled LTE-A network

uler that distributes evenly, in a TTI basis, the RBs among the active UEs, independently of the wireless channel conditions or QoS requirements.

3.2.1.2 SNR estimation and MCS selection

The UEs are located in various distances from the eNB. Assuming a fixed transmission power from eNB, the UEs experience different SNR levels. The influence of SNR heterogeneity is evident in the MCSs preferred for downlink transmissions. The better the LTE-A link quality, the higher order MCS is selected. As discussed in Section 2.2.2, in LTE-A networks, the MCS is determined according to CQIs that depict the downlink channel conditions and indicate the supported data rates [9]. Every CQI value corresponds to a specific MCS. The transmitted data are mapped into modulation symbols according to the MCSs supported by the LTE-A standard, e.g., QPSK and 64-QAM. The selected MCS affects the number of bits that can be carried per symbol.

For MCS selection, the SNR of each eNB-UE link must be estimated. In our study, we consider independent downlink channels with Rayleigh fading. Thus, the SNR is a random variable with average value γ and probability density function given by:

$$f(y) = \frac{1}{\gamma} e^{-\frac{y}{\gamma}} u(y), \tag{3.1}$$

where $u(y)$ is the unit step function.

3.2.2 D2D communication

The active UEs depicted in Fig. 5.1, i.e, UE_1 and UE_2, intend to initiate bidirectional communication among them. The packet arrival rate of each active UE depends on downlink data rate. After the reception of packets, the two UEs contend for Wi-Fi channel access using IEEE 802.11 DCF [27], and attempt to exchange data.

Erroneous packet transmissions might occur due to fluctuations of D2D links' quality. An active UE that fails to decode a packet asks for cooperation from idle UEs in close proximity that opportunistically overhear the packets exchanged during $UE_1 \leftrightarrow UE_2$ communication. The neighboring UEs decide whether they will join the relay set $\mathbf{R} = \{r_1, r_2, \ldots, r_N\}$, depending on their mode (transmission or idle), and whether NC packets can be transmitted during the cooperation.

In the channel model, fading is considered using the packet error rate (PER). The ergodicity of the fading process enables the use of bit error probability, which is directly related to PER [102]. The wireless channels between active UEs and relays are assumed to be independent of each other. We denote the PERs in the $UE_1 \leftrightarrow r_n$ and $UE_2 \leftrightarrow r_n$ D2D links, $r_n \in \mathbf{R}$, as $PER_{(UE_1 \leftrightarrow r_n)}$ and $PER_{(UE_2 \leftrightarrow r_n)}$, respectively.

The retransmissions of the packets of the active UE pair by the relays imply contention for channel access, which is resolved by the DCF method that uses various contention windows and backoff stages. A relay that is ready to transmit selects its backoff counter in a specific contention window range. Each relay may overhear zero, one or two packets of the two active UEs. In bidirectional communication, it is efficient that the relays serve simultaneously packets of both flows. However, the default DCF operation does not favor the selection of the relay with the higher number of overheard packets. To that end, the ACNC-MAC protocol can exploit the NC potential in cooperative D2D transmissions by prioritizing relays that are capable of performing encoded transmissions.

3.3 The ACNC-MAC protocol design

The ACNC-MAC protocol allows neighboring UEs to act as relays and perform cooperative transmissions, assisting the D2D communication of a UE pair. It goes beyond the SoA protocols by better exploiting NC opportunities arising in bidirectional D2D communication. The relays that overhear packets from both UEs transmit encoded packets, serving both flows at each communication round. ACNC-MAC prioritizes the transmissions of relays that can perform NC, maximizing the benefits of outband cooperative D2D communication. In this section, the ACNC-MAC protocol operation is detailed. It is compatible with the IEEE 802.11 standard and allows idle UEs within Wi-Fi range to act as relays. It adapts the relays' contention phase to the number of overheard packets, harnessing NC opportunities, and operates as a simple cooperative protocol when NC cannot be performed.

Upon the reception of a packet from the eNB, any of the two UEs can initiate a

communication round. Each UE that wishes to transmit contends for channel access by sensing the channel idle for DCF inter frame space (DIFS) and waits for a random backoff period. The cooperation of adjacent idle UEs is triggered by the transmission of a *request-for-cooperation* (RFC) frame, right after a short inter frame space (SIFS) waiting period. The RFC is sent by the active UE that fails to decode a packet transmitted by the other active UE. If it has a packet of its own to transmit, it is sent piggy-backed with the RFC, which initiates the cooperation phase of ACNC-MAC.

Once neighboring UEs receive the RFC, they decide whether they can act as relays. Each relay candidate receives at most two packets (one from each active UE), thus relays with zero, one or two packets may coexist in the relay set. ACNC-MAC prioritizes the relays with the highest number of overheard packets for the retransmission process, adopting a priority-based backoff counter selection mechanism. Letting i be the number of packets correctly decoded by a relay and $cw(k)$ the contention window of the k DCF backoff stage, each relay selects the backoff value with a contention window $cw_i \in [cw_{\min}, cw_{\max}]$ from the following ranges, as shown in Fig. 3.2:

$$cw_i \in \begin{cases} [2cw(k), 3cw(k) - 1], & \text{if } i = 0 \\ [cw(k), 2cw(k) - 1], & \text{if } i = 1 \\ [0, cw(k) - 1], & \text{if } i = 2 \end{cases} \tag{3.2}$$

For example, starting with cw_{min} equal to 32, the backoff values cw_2 of the relays that have received two packets will be chosen randomly in [0, 31]. If a collision occurs among relays, cw_{min} is doubled and the values are selected from the range [0, 63].

The relay that wins the contention transmits a special control frame, i.e., *eager-to-cooperate* (ETC), indicating the number of packets i to be sent (one packet or two packets encoded together). Transmitted after a SIFS period and a priority-based backoff period, ETC informs the two active UEs about the number of ACKs that will terminate the cooperation phase, and deters them from attempting new transmissions before all packets are delivered. It is possible that no ACK frames are transmitted, if none of the exchanged packets has been successfully decoded by any of the relays. Hence, the cooperation ends with the reception of an ETC frame, one ACK frame or two ACK frames by the UE pair.

The ACNC-MAC protocol can handle three different cases according to the number of packets delivered during the cooperation phase:

Case 0: No relay has correctly received any packet of the UE pair (Fig. 3.3(a)). No ACK frame is sent and the cooperation ends with the reception of an ETC frame. UE_1 wins the contention phase and transmits its packet p_1, which is not received by any relay, thus only the ETC is sent. In the meanwhile, during the first cooperation round, a packet q_1 has arrived in UE_2. Afterwards, UE_2 gains channel access and transmits q_1. UE_1 also has packet p_1 and sends RFC piggy-backed with p_1. The relays fail to receive either p_1 or q_1, so the cooperation ends with an ETC frame.

Case 1: Some relays have received only one packet while others have failed to decode any packet (Fig. 3.3(b)). The selected relay transmits ETC with one packet (of either

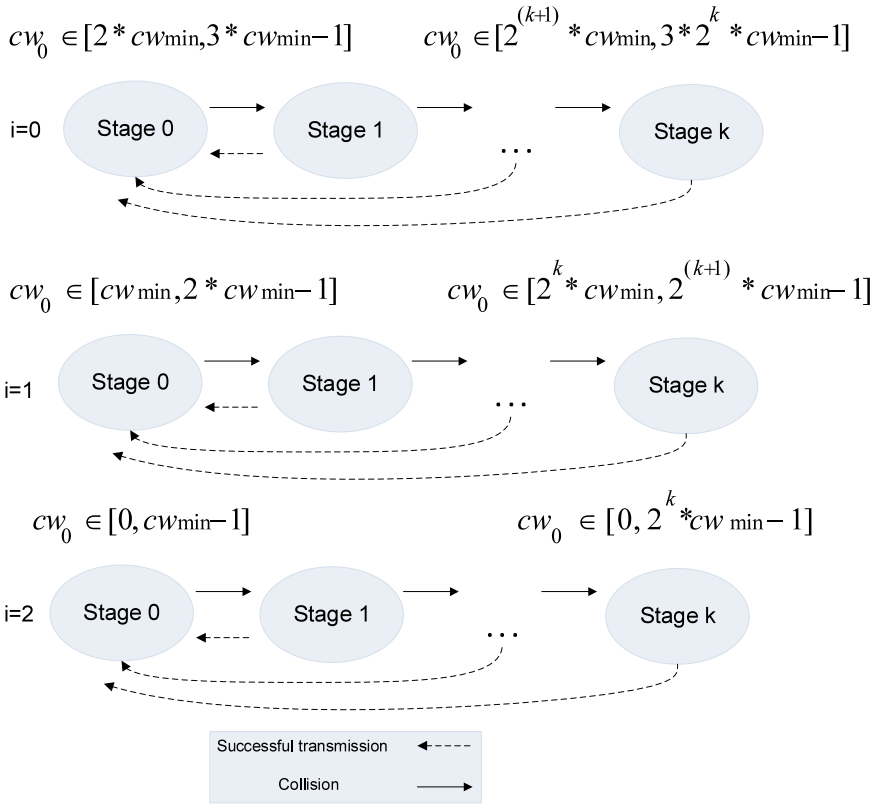
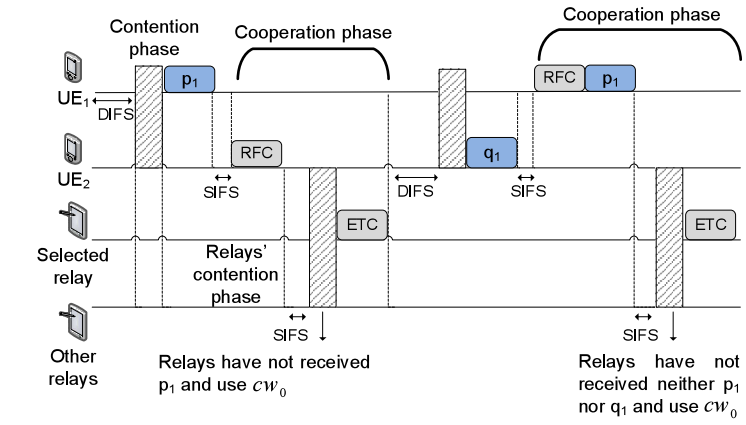


Figure 3.2: DCF backoff stages for ACNC-MAC

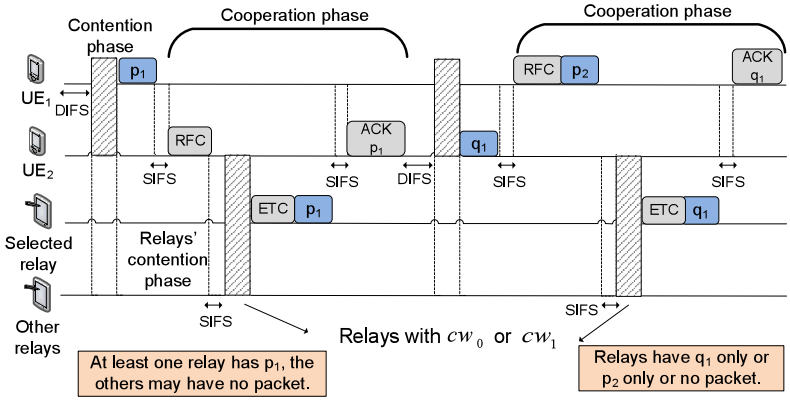
of the two UEs). ETC indicates that the cooperation phase will terminate by the transmission of only one ACK by the receiver UE. The packet p_1 of UE_1 is received by at least one relay, so the cooperation phase is terminated with the transmission of an ACK frame by UE_2. A packet q_1 has arrived in the buffer of UE_2, which wins the contention phase and sends q_1. UE_1 fails to decode it and asks for cooperation by sending RFC. As a new packet p_2 has arrived in buffer, UE_1 also sends p_2 with the RFC. Each relay has at most one packet (q_1 or p_2), thus relays with contention windows cw_0 or cw_1 may exist simultaneously. If a relay that received q_1 wins the contention phase, the cooperation ends when UE_1 transmits an ACK frame for q_1.

Case 2: This case occurs only when both UEs transmit packets and at least one relay receives them (Fig. 3.3(c)). The relay that wins the contention phase transmits the ETC piggy-backed with an NC packet. Hence, two ACK frames are expected to end the cooperation phase. UE_1 first sends its packet p_1 and UE_2 transmits its own packet q_1 with the RFC. Relays with zero, one or two packets may coexist and select their backoff periods using the corresponding contention windows cw_0, cw_1 and cw_2. The NC packet is transmitted by the relay that wins the contention phase along with ETC. At the end of cooperation phase, both UE_1 and UE_2 confirm the reception of q_1 and p_1, respectively.

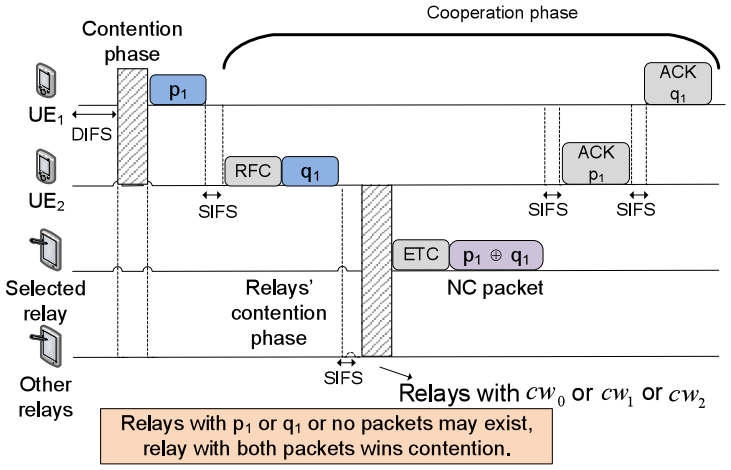
The NC operation is based on the XOR function and the butterfly structure [35]. In Fig. 3.4(a), the nodes S_1 and S_2 aim to send packets a_i and b_i to both D_1 and D_2, which can overhear the transmission of a_i and b_i, respectively. The relay R_2 sends an



(a) Case 0

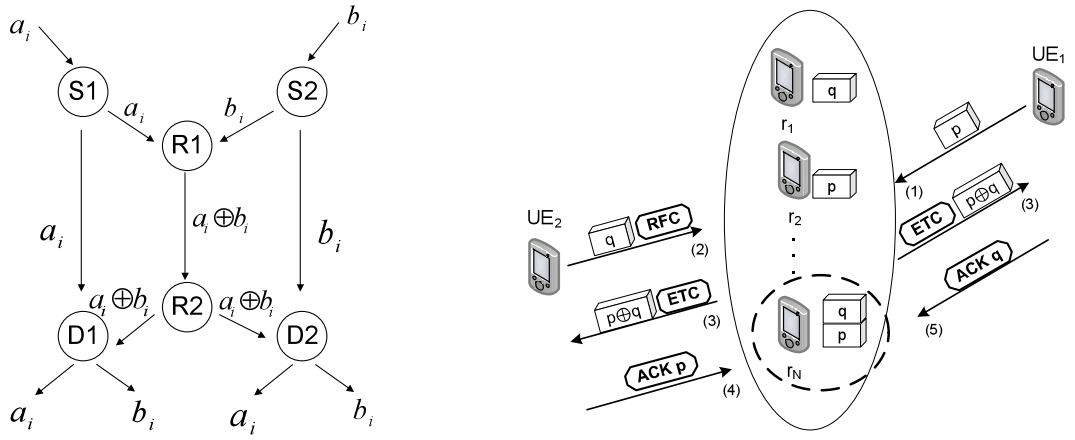


(b) Case 1



(c) Case 2

Figure 3.3: ACNC-MAC packet sequence for each case



(a) XOR function in the butterfly example (b) XOR function in ACNC-MAC protocol

Figure 3.4: The XOR function of network coding

encoded packet $a_i \oplus b_i$, thus delivering two packets with one transmission and achieving the multicast capacity. By receiving the encoded packet, the destination nodes D_1 and D_2 obtain the original packets b_i and a_i, respectively.

The NC in the ACNC-MAC protocol is performed in a similar manner. In Fig. 3.4(b)), UE_1 and UE_2 intend to exchange their packets p and q, respectively. The selected relay (r_N in the example) receives both packets, encodes them using the XOR function and multicasts the encoded packet $p \oplus q$, operating similarly as relay $R2$ in Fig. 3.4(a). Using the encoded packet $p \oplus q$ and its own packet, each UE decodes the original packet, e.g., UE_1 can decode q, using its own packet p.

3.4 Performance analysis

In this section, we present an analytical model for the achieved D2D network throughput using the ACNC-MAC protocol in saturated conditions. Furthermore, as the UEs that engage in D2D communication simultaneously receive the desired content from the eNB and share it with their peers, we also study the D2D MAC performance in the LTE–A context. Particularly, the packet exchange rate at D2D level is dictated by the packet arrival rates at the UEs, which are affected by i) the downlink resource scheduling policy, and ii) the cellular downlink channel conditions of the active UE pair. Considering the cross–network interactions between LTE–A and D2D communication levels described in Section 3.1, we also present an analytical model for the D2D throughput achieved by ACNC–MAC that captures the LTE–A parameters.

3.4.1 D2D throughput analysis in saturated conditions

We next present the analysis for the saturation throughput of ACNC-MAC. In saturated conditions, both sources transmit a packet at each round. The network throughput can

be defined as the ratio of the expected number of successfully delivered payload bits $\mathbb{E}[P]$ and the average time for a packet to be delivered to the destination $\overline{T}_{total}$:

$$\mathbb{E}[S] = \frac{\mathbb{E}[P]}{\mathbb{E}[T_{total}]}. \tag{3.3}$$

The average packet payload $\mathbb{E}[P]$ is a function of the probability P_1 that one packet is successfully delivered at the end of cooperation and the probability P_2 that packets of both sources are successfully received:

$$\mathbb{E}[P] = P_1\ell + 2P_2\ell, \tag{3.4}$$

where ℓ is the payload size. The total time required for the successful reception of two source packets is defined as:

$$\mathbb{E}[T_{total}] = \mathbb{E}[T_2]\,P_2 + \mathbb{E}[T_1]\,P_1 + \mathbb{E}[T_0]\,P_0. \tag{3.5}$$

The term $\mathbb{E}[T_{total}]$ is the weighted sum of three delay values that are related to cooperation phases with different outcomes.

The weights P_2, P_1 and P_0 are the probabilities of the different numbers of packets acknowledged at the end of cooperation and refer to the three cases mentioned in Section 3.3. P_2 is the probability that two packets are successfully received. This case occurs when at least one relay receives two packets and can encode them together. Letting $P_{e,(UE_1\leftrightarrow r)}$ and $P_{e,(UE_2\leftrightarrow r)}$ be the packet error probabilities in the D2D links $\forall r \in \mathbf{R}$, the probability that a relay r correctly receives both packets is given by:

$$P_{r,2} = (1 - P_{e,(UE_1\leftrightarrow r)})(1 - P_{e,(UE_2\leftrightarrow r)}). \tag{3.6}$$

If at least one relay decodes both source packets and can perform NC, the cooperation phase ends with the reception of two ACK frames. Therefore, the probability P_2 can be calculated as:

$$P_2 = 1 - \prod_{r=1}^{R}(1 - (1 - P_{r,2})). \tag{3.7}$$

The case that one packet is received by one of the two sources occurs with probability P_1. At least one of the relays receives one of the two source packets and the cooperation terminates with the reception of one ACK frame. The probability that a relay r correctly receives exactly one packet is given by:

$$P_{r,1} = (1 - P_{e,(UE_1\leftrightarrow r)}) + (1 - P_{e,(UE_2\leftrightarrow r)}) - 2(1 - P_{e,(UE_1\leftrightarrow r)})(1 - P_{e,(UE_2\leftrightarrow r)}). \tag{3.8}$$

Thus, P_1 can be derived by:

$$P_1 = \left[1 - \prod_{r=1}^{R}(1 - P_{r,1})\right] \prod_{r=1}^{R}(1 - P_{r,2}). \tag{3.9}$$

We denote as P_0 is the probability that no packet is received by any source finally. This event occurs when none of the relays receives any packets. As the probability that a relay r fails to receive both packets is equal to $P_{e,(UE_1 \leftrightarrow r)} P_{e,(UE_2 \leftrightarrow r)}$, the probability that all the relays do not receive any packet is the probability that no packet is acknowledged at the end of cooperation, which can be expressed as:

$$P_0 = \prod_{r=1}^{R} P_{e,(UE_1 \leftrightarrow r)} \prod_{r=1}^{R} P_{e,(UE_2 \leftrightarrow r)}. \tag{3.10}$$

The aforementioned probabilities are associated with the delay values $\mathbb{E}[T_2]$, $\mathbb{E}[T_1]$ and $\mathbb{E}[T_0]$. $\mathbb{E}[T_2]$ is the average time required for the successful reception of two packets and $\mathbb{E}[T_1]$ is the average time required for the successful reception of only one packet. The term $\mathbb{E}[T_0]$ is the average delay of a cooperation phase that does not deliver any packet, since the relays have failed to receive any of the transmitted packets. Each of these terms comprises of two components: i) the minimum average delay in case of perfect synchronization of relays (contention-free cooperation phase), and ii) the additional delay induced by the contention of the relays during the cooperation phase. These values differentiate according to the number of packets the relay transmits. Under these considerations, the average delay induced in a cooperation phase that ends with i ACK frames, $i \in \{0, 1, 2\}$, is:

$$\mathbb{E}[T_i] = \mathbb{E}[T_{i,min}] + \mathbb{E}[T_{i,cont}]. \tag{3.11}$$

When no packet is acknowledged, namely $i = 0$, the minimum average delay is:

$$\begin{aligned} \mathbb{E}[T_{0,min}] = {} & DIFS + T_{p_1} + T_{RFC} + T_{p_1'} + SIFS \\ & + T_{ETC} + 2SIFS + \mathbb{E}[r](SIFS + T_{ETC}). \end{aligned} \tag{3.12}$$

Similarly, for the case that one packet only is acknowledged, namely $i = 1$, the minimum average delay of contention-free cooperation phase is:

$$\begin{aligned} \mathbb{E}[T_{1,min}] = {} & DIFS + T_{p_1} + T_{RFC} + T_{p_1'} + SIFS \\ & + T_{ETC} + 2SIFS + T_{ACK} + \mathbb{E}[r](SIFS + T_{ETC} + T_{p_1}). \end{aligned} \tag{3.13}$$

When both sources receive their desired packets the minimum average delay of the cooperation phase is equal to:

$$\begin{aligned} \mathbb{E}[T_{2,min}] = {} & DIFS + T_{p_1} + T_{RFC} + T_{p_1'} + SIFS + T_{ETC} \\ & + 2SIFS + 2T_{ACK} + \mathbb{E}[r](SIFS + T_{ETC} + T_{p_1 \oplus p_1'}). \end{aligned} \tag{3.14}$$

The average delay of a cooperation phase includes also the term $\mathbb{E}[T_{i,cont}]$, which refers to the delay due to relays contention and is expressed as:

$$\mathbb{E}[T_{i,cont}] = \mathbb{E}[r]\,\mathbb{E}[T_{c,i}], \ i \in \{0, 1, 2\}, \tag{3.15}$$

where $\mathbb{E}[r]$ is the expected number of retransmissions, directly related with $PER_{(UE_1 \leftrightarrow r)}$ and $PER_{(UE_2 \leftrightarrow r)}$ [103]. The term $\mathbb{E}[T_{c,i}]$ corresponds to the average time required to

transmit packets during the contention phase among the relays and obtains a different value for each i, given that the number of packets a relay receives varies. Furthermore, the average backoff times selected by the relays from different ranges, according to the number of packets i they wish to transmit, changes as well. They can be estimated using the backoff counter model described in [40].

3.4.2 Cross-network D2D throughput analysis

In this section, we provide a cross-network theoretical model of the throughput performance of the ACNC-MAC protocol, used for D2D data exchange between two UEs that concurrently receive packets from cellular links. The proposed model jointly captures the dynamics of both cellular and D2D connectivity.

As already explained, the ACNC-MAC cooperation terminates with one of three possible outcomes (ACNC-MAC cases), according to the number of packets originally transmitted (one or two) and the number of packets successfully delivered (up to two). Given that the duration of each communication round varies analogously, the delay induced by each outcome must be weighted by the corresponding probability. The probability of occurrence of a case consists of two factors: i) the probability that a packet has arrived to either one or both active UEs, i.e., *packet arrival probability*, and ii) the probability that zero, one or two packets are acknowledged at the end of cooperation, i.e., *packet reception probability*. Therefore, we formulate the packet arrival and packet reception probabilities for each case.

As ACNC-MAC employs the IEEE 802.11 DCF, the channel access of the UEs that particate in the D2D data exchange must be modeled. If the D2D network operates in saturation, i.e., the UEs always have packets to transmit, the bi-dimensional Bianchi model [104] is utilized. It employs a Markov chain to model the backoff window size, used for the estimation of the steady state transmission and collision probabilities required for the throughput estimation. In case of non-saturated conditions, the Malone model [105] is employed, which introduces the idle state in the Markov chain, capturing the event that a UE remains idle between two packet arrivals.

The considered D2D network is formed of two sets of UEs, i.e., the active UE pair and the idle UEs (relays), which operate under different traffic conditions. The cellular link dependent packet arrival rates impose that the buffer of an active UE might be empty, i.e., it operates in non-saturated conditions. Hence, for the modeling of the backoff counter of the active UEs, we use the Malone model [105]. In contrast, it can be observed that the relays operate in saturated conditions, as they always transmit at least the ETC frame. All relays participate in the contention phase, but only the relays that have received the most packets are considered to be *active* and may experience collisions. The channel access of the relays can be modeled by the Bianchi model [104] using different number of active relays for each ACNC-MAC case. Therefore, the active relay set size per case must be analytically derived.

3.4.2.1 Packet arrival probabilities

The D2D network operates in conjunction with the cellular network, thus the packet arrival rate is regulated by the eNB that serves the active UEs[1]. The downlink data rate is affected by parameters related to the LTE-A network setup and the wireless channel conditions of the cellular links. As already mentioned, the eNB employs a scheduling algorithm that distributes the RBs to UEs. Moreover, the downlink channel state effect is apparent as each UE declares the MCS it supports according to the downlink SNR values. This process might cause variations to the throughput achieved for the UE. More specifically, the packet arrival rate is affected by the number K of concurrently active UEs, the number N_{RB} of available RBs, the packet size ℓ, the packet scheduling policy and the MCS choices. Considering that S different MCSs are available, the packet arrival rate at a UE can be estimated as:

$$\lambda = \sum_{i=1}^{S} \pi_i \frac{L\Big(MCS = i, \lfloor \mathbb{E}\left[b\right] \rfloor\Big)}{TTI \cdot \ell}, \tag{3.16}$$

where the TBS $L(MCS = i, \lfloor \mathbb{E}\left[b\right] \rfloor)$ can be found in [106]. The expected number $\mathbb{E}\left[b\right]$ of allocated RBs per UE depends on the scheduling policy. The probability π_i that the ith MCS is selected is derived as:

$$\pi_i = \int_{\gamma_{thr}^{(i)}}^{\gamma_{thr}^{(i+1)}} f(y)dy = e^{\frac{\gamma_{thr}^{(i+1)}}{\gamma}} - e^{\frac{\gamma_{thr}^{(i)}}{\gamma}}, \tag{3.17}$$

where γ is the average SNR and $[\gamma_{thr}^{(i)}, \gamma_{thr}^{(i+1)}]$ denotes the SNR range that corresponds to the MCS i.

For the throughput analysis, the offered load of the active UE pair can be modeled using the Poisson packet arrival process with mean value λ (packets/s). Particularly, in our model, we consider two active UEs with corresponding packet arrival rates λ_1 and λ_2. Once a packet transmitted by the eNB is received, the UE joins the contention phase following the IEEE 802.11 DCF rules. The two active UEs are not in saturated conditions, as the packets from eNB arrive in variable intervals. For the formulation of probabilities of the ACNC-MAC cases, we consider the probabilities that j packets arrive at the active UEs. Given that at least one packet is required to initiate the D2D communication, we define the probability $P(D_j)$, $j \in \{1, 2\}$ that j packets arrive at the UE pair.

Lemma 1. *A packet arrives in both UEs with probability:*

$$P(D_2) = (1 - \mathrm{e}^{-\lambda_1 \mathbb{E}[T_{slot}]})(1 - \mathrm{e}^{-\lambda_2 \mathbb{E}[T_{slot}]}), \tag{3.18}$$

where $\mathbb{E}\left[T_{slot}\right]$ is the time spent at each state of the Markov chain considering the Malone model [105].

Proof. The proof of Lemma 1 is provided in Appendix 3.A.1. ∎

[1]Our model is also applicable in case that the UEs belong to different cells.

Lemma 2. *Exactly one packet arrives at the D2D network, i.e., only one of the two active UEs receives a packet from the eNB, with probability:*

$$P(D_1) = (1 - \mathrm{e}^{-\lambda_1 \mathbb{E}[T_{slot}]}) + (1 - \mathrm{e}^{-\lambda_2 \mathbb{E}[T_{slot}]}) \\ -(1 - \mathrm{e}^{-\lambda_1 \mathbb{E}[T_{slot}]})(1 - \mathrm{e}^{-\lambda_2 \mathbb{E}[T_{slot}]}). \tag{3.19}$$

Proof. When the contingency D_1 occurs, a packet arrives at either of the UEs but not in both of them simultaneously. In a similar manner as in Lemma 1, the addition rule is used for the estimation of $P(D_1)$. ∎

3.4.2.2 Packet reception probabilities

The end of cooperation phase is indicated by the reception of i) an ETC frame, if no packet has been successfully received by any relay, ii) a single ACK frame, if at least one relay decodes exactly one packet and no relay has two packets, or iii) two ACK frames, if at least one relay receives packets from both active UEs and performs NC. Each case ensues from the different number of data packets overheard by the $|\mathbf{R}| = N$ idle UEs. It can be observed that the contingencies of zero (C_0), one (C_1) or two ACK frames (C_2) are mutually exclusive. Furthermore, the contingency D_1 of packet arrival in only one UE and the contingency D_2 of packet arrival in both UEs concurrently form a partition of sample space $\mathcal{D}$, as $D_1 \cap D_2 = \emptyset$ and $D_1 \cup D_2 = \mathcal{D}$. It should be also noted that each of the events C_i, $i \in \{0, 1, 2\}$ that form the sample space $\mathcal{C}$ occur after the packet arrival events $D_j \in \mathcal{D}$, $j \in \{1, 2\}$.

Lemma 3. *If event C_i occurs after event D_j with conditional probability $P(C_i|D_j)$, the probability that C_i occurs is:*

$$P(C_i) = P(C_i|D_1)P(D_1) + P(C_i|D_2)P(D_2), i \in \{0, 1, 2\}. \tag{3.20}$$

Proof. For the events $D_j \in \mathcal{D}$, it holds that $P(D_j) > 0, j \in \{1, 2\}$. Then, for any event $C_i, i \in \{0, 1, 2\}$, $P(C_i)$ can be calculated using the total probability formula as $P(C_i) = \sum_j P(C_i \cap D_j) = \sum_j P(C_i|D_j)P(D_j)$. ∎

We next define $H_{i,j}$ as the event of termination of cooperation with i ACK frames, i.e., the event that the relays have i packets, after the transmission of j packets, and $P(H_{i,j}) \equiv P(C_i|D_j)$ as its corresponding probability. The duration of each transmission round varies with the number of packets exchanged. Hence, the total time required for the packet(s) successful delivery, or the end of cooperation with ETC frame is weighted using the following probabilistic coefficients:

(i) *Cooperation ends with ETC frame* (C_0): Either one or both UEs transmit a packet. The UE that wins the contention phase transmits its packet and the other UE transmits its own packet (if any) piggy-backed with the RFC frame. This case occurs with probability:

$$P(C_0) = P(H_{0,1})P(D_1) + P(H_{0,2})P(D_2), \tag{3.21}$$

where

$$P(H_{0,j}) = \prod_{n=1}^{N} \left[PER_{(UE_1 \leftrightarrow r_n)} P(D_j)\right] + \prod_{n=1}^{N} \left[PER_{(UE_2 \leftrightarrow r_n)} P(D_j)\right], j \in \{1,2\}. \tag{3.22}$$

This probability corresponds to the case that none of the relays succeeds in receiving any packet from the UE pair.

(ii) *Cooperation ends with one ACK frame* (C_1): One or two packets are sent and the relays receive one of them. If both UEs send a packet, all relays fail to correctly decode both packets. The corresponding probability is:

$$P(C_1) = P(H_{1,1})P(D_1) + P(H_{1,2})P(D_2), \tag{3.23}$$

where the probability that at least one relay has exactly one packet is:

$$P(H_{1,j}) = 1 - \prod_{n=1}^{N} (1 - P(H_{1,j}^{(n)})), j \in \{1,2\}. \tag{3.24}$$

One or two packets are sent and some relays overhear one packet. If two packets are sent, no relay receives both packets. The probability of reception of exactly one packet by relay r_n when only one UE has transmitted is:

$$P(H_{1,1}^{(n)}) = (1 - PER_{(UE_1 \leftrightarrow r_n)})P(D_1) + (1 - PER_{(UE_2 \leftrightarrow r_n)})P(D_1), \tag{3.25}$$

and when both UEs have transmitted packets, it is:

$$P(H_{1,2}^{(n)}) = (PER_{(UE_1 \leftrightarrow r_n)} + PER_{(UE_2 \leftrightarrow r_n)} - 2PER_{(UE_1 \leftrightarrow r_n)} PER_{(UE_2 \leftrightarrow r_n)})P(D_2). \tag{3.26}$$

(iii) *Cooperation ends with two ACK frames* (C_2): This case might occur when both UEs have transmitted packets and at least one relay receives both of them. Thus, the probability that an NC packet is transmitted is:

$$P(C_2) = P(H_{2,2})P(D_2), \tag{3.27}$$

with

$$P(H_{2,2}) = 1 - \prod_{n=1}^{N} (1 - P(H_{2,2}^{(n)})) \tag{3.28}$$

and

$$P(H_{2,2}^{(n)}) = (1 - PER_{(UE_1 \leftrightarrow r_n)})(1 - PER_{(UE_1 \leftrightarrow r_n)})P(D_2), \tag{3.29}$$

which is the probability that a given relay overhears both packets.

As the duration of cooperation phase depends on the number of transmitted packets by the UE pair and the number of packets overheard by the relays, the aforementioned probabilities are used for the throughput estimation in Section 3.4.2.4.

3.4.2.3 Estimation of the active relay set size

We thereupon estimate the number of relays that are active during cooperation. Collisions occur only among relays with the highest number of packets, which gain the highest priority in backoff selection.

Definition 1. *For each ACNC-MAC case $i \in \{0, 1, 2\}$, we define the set of relays whose transmissions may lead to collisions as* active relay set $M_i \subseteq R$ *with expected size* $|M_i|$.

For the estimation of $|M_i|$, two probabilistic coefficients must be calculated for each case i: i) the probability $P(H_i)$ that at least one relay has received i packets, and ii) the probability $P(|M_i| = k)$ that k relays have received i packets.

Lemma 4. *Letting k be the number of relays that have zero, one or two packets in each ACNC-MAC case respectively, the expected active relay set size $|M_i|$ is expressed as:*

$$|M_i| = \sum_{k=1}^{N} kP(|M_i| = k), i \in \{0, 1, 2\}, \tag{3.30}$$

where the probability that $|M_i| = k$ is given by:

$$P(|M_i| = k) = \binom{N}{k} P(H_i)^k (1 - P(H_i))^{N-k}. \tag{3.31}$$

Proof. The proof of Lemma 4 is provided in Appendix 3.A.2. ∎

3.4.2.4 Throughput analytical formulation

Having presented the essential components for modeling the throughput performance of the ACNC-MAC protocol , we next provide the throughput analysis. For the throughput estimation, the expected duration of a D2D communication round $\mathbb{E}[T_{i,j}]$, with $i \in \{0, 1, 2\}$ and $j \in \{1, 2\}$ must be derived.

Lemma 5. *The value of $\mathbb{E}[T_{i,j}]$ is estimated as follows:*

$$\begin{aligned} \mathbb{E}[T_{i,j}] = \overbrace{\underbrace{\mathbb{E}[T_{slot}] + SIFS + T_{ETC}}_{\mathbb{E}[T_{init}]} + \mathbb{E}[r]\, x_{i,j} + y_{i,j}}^{\mathbb{E}\left[T_{i,j}^{min}\right]} \\ + \underbrace{\mathbb{E}[r]\,\mathbb{E}[Tc_i]}_{\mathbb{E}[T_i^{cont}]}, \end{aligned} \tag{3.32}$$

where $\mathbb{E}[T_{i,j}]$ consists of two components: $\mathbb{E}\left[T_{i,j}^{min}\right]$ is the minimum duration of a contention-free cooperation phase and $\mathbb{E}[T_i^{cont}]$ is the delay due to the relays' contention.

Proof. The proof of Lemma 5 is provided in Appendix 3.A.3. ∎

Proposition 1. *The expected ACNC-MAC throughput is given by Eq. (3.33), where* $\mathbb{E}[P]$ *is the average correctly received useful bits,* $\mathbb{E}[T_{total}]$ *is the average time required for a packet to be delivered to its destination and* ℓ *the packet payload size.*

$$\mathbb{E}[S] = \frac{\overbrace{\ell(P(H_{1,1}) + P(H_{1,2})) + 2\ell P(H_{2,2})}^{\mathbb{E}[P]}}{\underbrace{\begin{array}{c} P(H_{0,1})\mathbb{E}[T_{0,1}] + P(H_{0,2})\mathbb{E}[T_{0,2}] + P(H_{1,1})\mathbb{E}[T_{1,1}] \\ + P(H_{1,2})\mathbb{E}[T_{1,2}] + P(H_{2,2})\mathbb{E}[T_{2,2}] \end{array}}_{\mathbb{E}[T_{total}]}} \tag{3.33}$$

The terms $\mathbb{E}[T_{i,j}]$ given by Eq. (3.32) and the probabilistic coefficients given by Eqs. (3.22), (3.24) and (3.28) are used for the throughput estimation. $\mathbb{E}[P]$ is weighted by the probabilities that one or two packets are successfully delivered. The delay values that constitute the average delay term are weighted by the probabilities $P(H_{i,j})$, $i \in \{0, 1, 2\}$ and $j \in \{1, 2\}$, i.e., the probabilities of each of the five possible outcomes inferred by the conjunction of the packet arrival contingencies D_1 and D_2 and the ACNC-MAC cases C_0, C_1 and C_2.

3.5 Model validation and performance assessment

In this section, we evaluate the analytical models presented in Section 3.4 and we assess the performance of the proposed protocol in saturated and non-saturated network traffic conditions, in comparison with the most related SoA, i.e, the NCCARQ-MAC protocol [40]. We also thoroughly study the ACNC-MAC performance for different downlink packet scheduling policies, MCSs and numbers of active UEs. Moreover, we present the performance results for video transmission scenarios and investigate the influence of different idle UE deployments. In our study, we consider three different network cases, i.e., i) a D2D network that operates in saturated conditions (case A), ii) a D2D network in non-saturated operation (case B), and iii) a D2D network that resides in an LTE-A cell and operates under the impact of cellular network parameters (case C).

In the first part of the performance evaluation, elabored in Section 3.5.2, we validate the proposed analytical models and provide comparative results of the ACNC-MAC and NCCARQ-MAC protocols [40]. More specifically, the case A (Section 3.5.2.1) serves as the network paradigm that validates the saturation throughput analysis presented in Section 3.4.1. In Section 3.5.2.2, the performance results of case B using various traffic load levels are presented. In Section 3.5.2.3, the cross-network D2D throughput analysis presented in Section 3.4.2 is validated using the set-up of case C. For cases B and C, we compare ACNC-MAC with a modified version of NCCARQ-MAC that permits the protocol application in non-saturated conditions incited by D2D communication. With NCCARQ-MAC, the relays cooperate only when they receive packets from both UEs and can perform NC transmissions.

The second part of the performance evaluation presented in Section 3.5.3 is a thorough experimental evaluation of ACNC-MAC considering the network case C. In detail, we study the effect of various LTE-A network parameters, i.e., selection of MCSs (Section 3.5.3.1) and downlink packet scheduling policies (Section 3.5.3.2), on the ACNC-MAC performance, and the influence of the distributions of the idle UEs (relay candidates) in a video transmission scenario (Section 3.5.3.3).

In our simulations, we use a C++ integrated simulator that implements the different downlink packet scheduling policies and applies the ACNC-MAC protocol rules. The simulation setup and the metrics considered for the performance evaluation are analytically described in Section 3.5.1.

3.5.1 Simulation setup and evaluation metrics

In all network cases, for the D2D links, we use PER as channel quality indicator, as described in Section 3.2.2. We also assume that N relays assist the UE pair's communication and all D2D links experience the same PER[2]. The rest of the simulation parameters are summarized in Table 3.1.

In the network cases A and B, we consider the topology of Fig. 5.1 and simulate the bidirectional communication of the two active UEs, UE_1 and UE_2, aided by $N = 5$ adjacent and initially idle UEs that can be used as relays. It is assumed that $PER_{(UE_1 \leftrightarrow r)} = PER_{(UE_2 \leftrightarrow r)}$ and . We also assume that $PER_{(UE_1 \leftrightarrow UE_2)} = 1$, thus, a cooperation phase is always initiated. In case A, the UEs always have packets to transmit (saturated conditions), whereas in case B, the UEs generate packets according to a Poisson traffic model with the same intensity λ. Two data rate scenarios are tested, using the D2D network parameters shown in Table 3.1. The transmission rates for the active UEs are $R_{s,r} = \{6, 54\}$ Mb/s for low and high data rate scenario, respectively, while the relays transmit in both scenarios at a constant rate $R_{r,s} = 54$ Mb/s.

In the network case C, we consider that the UE pair of Fig. 5.1 receives data from the eNB, which serves a total of K UEs in the cell. The UEs belong to either of two SNR classes m_{high} and m_{low} of high and low SNR, respectively and each class includes $K/2$ UEs. We set a threshold SNR, SNR_{thres}, as a bound between the two classes. All UEs that experience SNR values higher than SNR_{thres} use 64-QAM and belong to the m_{high} class, while the rest of them use QPSK or 16-QAM and belong to the m_{low} class. For UEs with the same modulation scheme, different coding rates may be used. The minimum SNR value derived in the simulations corresponds to the lowest SNR threshold for the MCS with the lowest modulation order and coding rate. In LTE-A transmissions, the Round Robin scheduler is used, unless otherwise stated. The active UEs have both their LTE-A and Wi-Fi interfaces concurrently active, whereas the relays use only the Wi-Fi connection. The energy consumption of the active UEs, denoted as E, is the sum of the average energy consumed during the data reception from the cellular link, denoted as $\mathbb{E}[E_{LTE-A}]$ and

[2]We use a fixed PER, since different PER values affect the performance as expected, without influencing our conclusions.

Table 3.1: Simulation parameters for model validation and performance evaluation of ACNC-MAC protocol

Parameter	**Value**
Cellular network (case C)	
N_{RB}	100
Bandwidth	20 MHz
Modulation schemes	QPSK, 16-QAM, 64-QAM
Channel model	Rayleigh fading
SNR classes	low:{QPSK, 16-QAM}, high:{64-QAM}
TTI	1 ms
P_{Rx}^{LTE-A}	4 W [107]
D2D network	
Data Tx rate (Mb/s)	active UEs: 54 (case C), relays: 54 (all cases)
Control frame Tx rate (Mb/s)	6
Payload size	1500 bytes
ETC	16 bytes
DIFS	50 μs
SIFS	10 μs
RFC, ACK	14 bytes
PHY header	96 μs
$cw_{\min}$	32
$P_{Rx} = P_{idle}$, P_{Tx} (mW)	1340, 1900 [108]
PER	[0-0.9] (cases A and B), 0.2 (case C)
λ (case B)	[100-2500]
UE characteristics	$C_0 = 1300$ mAh, $V_0 = 3.7$ V

the energy consumed in D2D transmissions using ACNC-MAC, denoted as $\mathbb{E}[E_{D2D}]$. The LTE-A interface of the relays is not active and only the energy consumption in Wi-Fi interface is considered, thus we set $E = \mathbb{E}[E_{D2D}]$.

In the simulation scenarios referring to case C (Sections 3.5.2.3, 3.5.3.1 and 3.5.3.2), the UE pair uses the ACNC-MAC protocol in order to exchange files of 5 MB size concurrently downloaded through the cellular links. Furthermore, considering the escalating proliferation of multimedia-based mobile applications, we assess the ACNC-MAC performance in video exchange scenarios (Section 3.5.3.3), where a video sequence is transmitted by the eNB to the UEs and is further exchanged by the UE pair. The video data are delivered by the eNB in H.264/SVC video compression format [109]. The JSVM 9.19 software [110] is used for the encoding of the "BUS" QCIF video sequence with frame rate 15 frames/sec. The generated packets are transmitted over the LTE-A link and once they are received by the UEs, their transmission with ACNC-MAC is initiated.

Regarding the metrics used for the performance evaluation, we should mention that the ACNC-MAC performance is evaluated in terms of aggregated D2D network throughput and energy efficiency, i.e., the amount of payload bits exchanged over the total energy consumption (measured in bits/Joule) [111]. The amount of useful bits received is the sum of useful bits received by the final destinations, i.e., the sum of bits of useful data received by the D2D pair. The total energy consumption refers to the energy consumed

by the D2D pair and the relays. Additionally, aiming to gain a better insight on the induced energy consumption, in the simulation scenario of Section 3.5.3.3, we estimate the average battery drain ΔC (mAh) of the UE pair and the relays as follows [112]:

$$\Delta C = C_0 - C, \tag{3.34}$$

where C_0 is the initial battery capacity. The value C is the expected battery capacity that can be calculated as:

$$C = (E_0 - E)/(V_0 \cdot 60^2), \tag{3.35}$$

where V_0 is the battery voltage, $E_0 = V_0 \cdot C_0 \cdot 60^2$ is the initial energy and E is the total energy consumption of each UE measured in Joules.

3.5.2 Analysis validation and comparison with NCCARQ-MAC

We thereupon validate the proposed analytical models (Sections 3.5.2.1 and 3.5.2.3) and compare the performance of the ACNC-MAC and NCCARQ-MAC protocols.

3.5.2.1 Saturation throughput analysis validation and performance results for case A

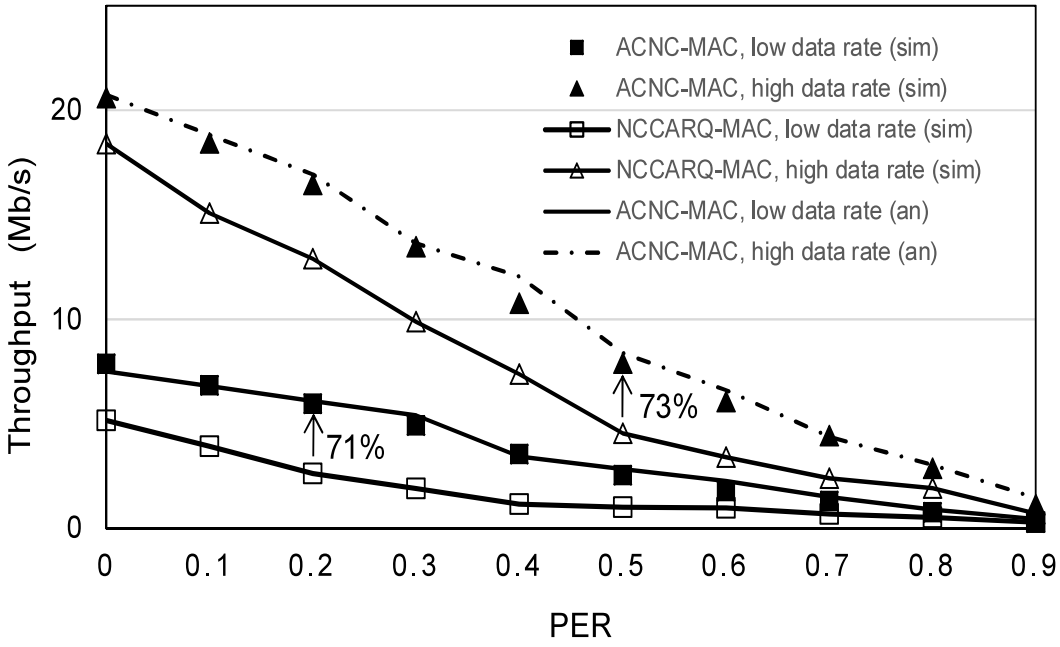
Figure 3.5(a) shows the throughput performance for the case A with regard to different PERs considering two different data rate scenarios. As observed, the simulation and theoretical results for throughput performance match, thus verifying the proposed throughput analysis for saturated conditions. ACNC-MAC achieves better performance than NCCARQ-MAC, as it better exploits cooperation opportunities by serving at least one packet per communication round. For PERs in [0, 0.5], ACNC-MAC achieves an improvement up to 71% and 73% in low and high data rate scenario, respectively.

In Fig. 3.5(b), the energy performance in the case A is depicted. It is obvious that the energy efficiency curves are similar to throughput curves, as expected. However, as PER increases, more retransmissions are required in order to correctly deliver each packet, thus the energy efficiency for each successful packet transmission reduces. Still, the ACNC-MAC protocol performs better than NCCARQ-MAC in both data rate scenarios for all PER values. This can be justified by the fact that more useful bits are delivered under the same energy consumption, as ACNC-MAC allows for relays retransmissions, even when NC is not possible. In contrast, in each cooperation round of NCCARQ-MAC protocol, either two packets are delivered or none at all. Notably, the gain of ACNC-MAC is higher when high data rates are used, reaching a 71% increase for PER=0.3.

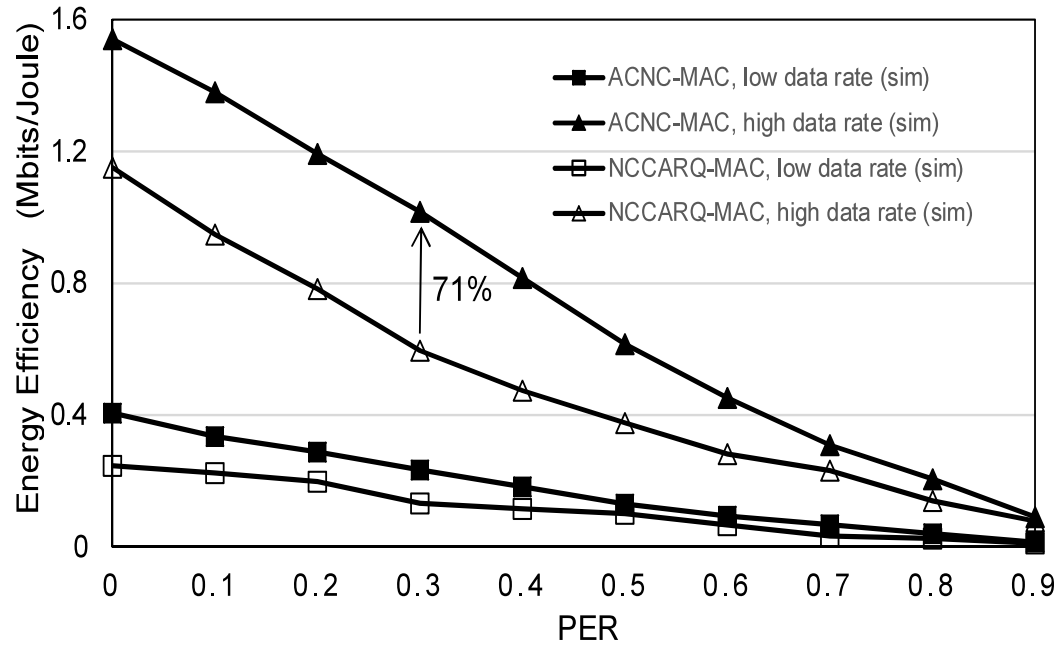
3.5.2.2 Performance results for case B

We thereupon assess the performance of ACNC-MAC and NCCARQ-MAC protocols considering a D2D network that operates under non-saturated conditions.

First, in Fig. 3.6(a), the throughput performance for the case B is illustrated. For lower traffic intensity, namely $\lambda < 300$, the gains of NC are not fully exploited, due



(a) Total throughput in saturation



(b) Energy efficiency in saturation

Figure 3.5: ACNC-MAC performance results in saturated conditions (case A)

to scarce packet arrivals. Instead, as traffic in the active UEs increases, NC possibility becomes higher, leading to a throughput increase. It can be seen that ACNC-MAC achieves throughput gains up to 41% in low rate scenario and up to 38% in high rate scenario, for λ in the range [900, 2300].

Continuing, Fig. 3.6(b) shows the performance in terms of energy efficiency for the case B. Notably, the energy efficiency achieved by the ACNC-MAC protocol is higher than NCCARQ-MAC in both data rate scenarios. For high data rate in particular, the energy energy efficiency is $32 - 39\%$ higher, when ACNC-MAC is used. Regarding the low rate scenario, we can observe that the resulting energy efficiency of both protocols deteriorates. However, it should be noted that with the ACNC-MAC protocol the energy efficiency is almost doubled comparing to NCCARQ-MAC.

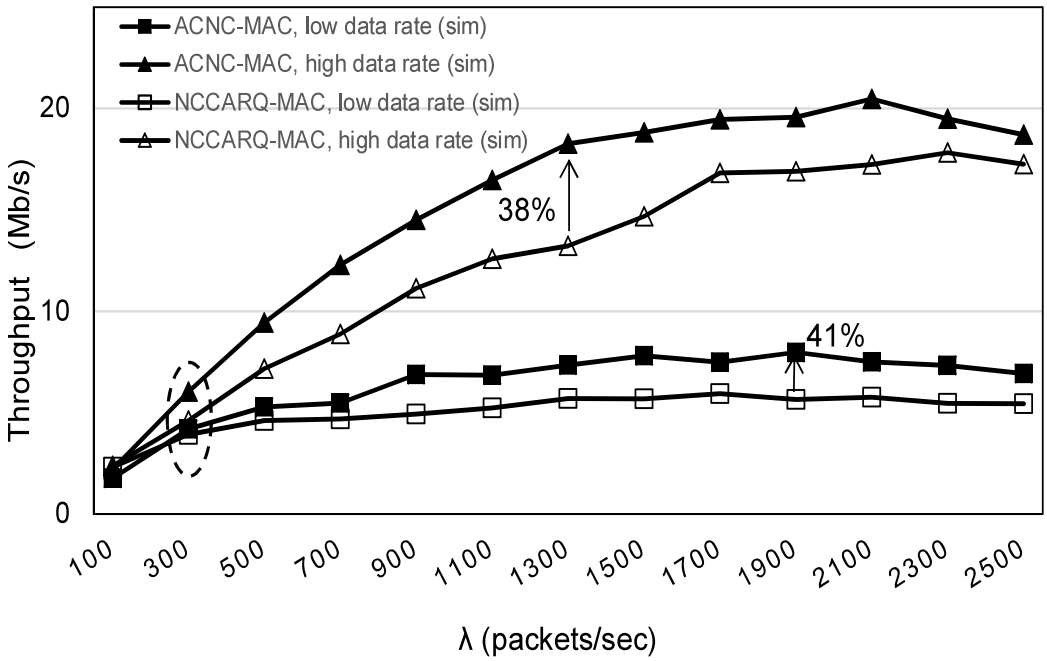
It is also worth pointing out that throughput and energy efficiency plots exhibit a similar behavior in case of saturated conditions, whereas they differentiate when varied traffic values are used. Moreover, as traffic intensity increases, the energy efficiency remains at the same levels. These observations can be explained by the fact that, for small λ values, fewer packets are delivered and more idle slots exist. Also, when higher λ values are used, more packets are delivered. However, the energy efficiency is similar, since less idle slots exist but more packet receptions occur, whereas P_R is equal to P_I.

3.5.2.3 Cross-network analysis validation and performance results for case C

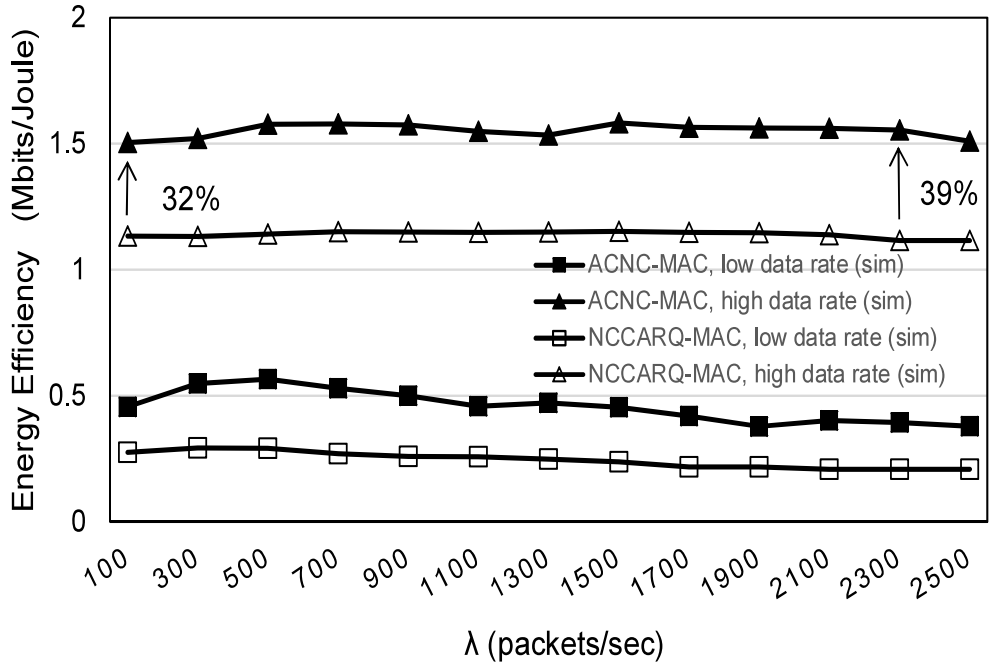
For the validation of the throughput analysis that will be presented next, we assume a cell with $K \in \{20, 40, 60, 80\}$ active UEs, where the number of idle UEs that can be used as relays is equal to $N = 5$.

First of all, we should note that ahe match of theoretical and simulation results corroborate the throughput analysis, as shown in Fig. 3.7. Moreover, it can be observed that the ACNC-MAC protocol outperforms NCCARQ-MAC in terms of throughput, as it can exploit more efficiently the cooperation opportunities. More specifically, the ACNC-MAC throughput is 134% and 226% higher for the m_{low}-m_{high} and m_{high}-m_{high} UE pair, respectively ($K = 80$). Notably, when the cell congestion, i.e., the value K, increases, the throughput achieved by both of the protocols under comparison deteriorates. As more UEs are served in each TTI, fewer RBs are allocated to each UE, reducing the downlink data rate. Thus, the packet arrival rates also reduce, increasing the duration of data exchange between the UE pair, as more communication rounds are required to deliver the same amount of data. Comparing the ACNC-MAC throughput for $K = 20$ and $K = 80$, we observe a decrease of 62% for m_{high} UEs and 67% for m_{low}-m_{high} UEs. However, the ACNC-MAC throughput remains higher than the NCCARQ-MAC throughput. The gain increases along with K, as packet arrival rates decrease, reducing the NC opportunities. Hence, fewer fruitful communication rounds occur with NCCARQ-MAC.

It can be also seen that ACNC-MAC achieves higher energy efficiency than NCCARQ-MAC in all scenarios (Fig. 3.8). More transmission rounds fail to deliver packets when NCCARQ-MAC is used, whereas ACNC-MAC allows retransmissions by relays, even when



(a) Total throughput for various λ and PER=0



(b) Energy efficiency for various λ and PER=0

Figure 3.6: ACNC-MAC performance results in non-saturated conditions (case b)

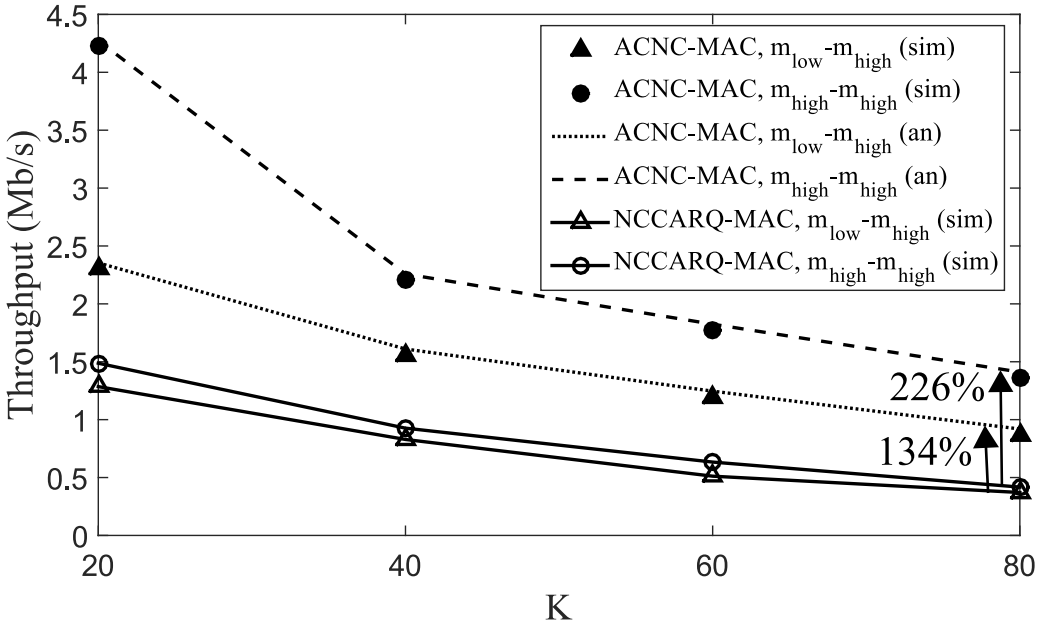


Figure 3.7: D2D throughput for different SNR classes vs. K

only one packet exists in at least one of them. For this reason, ACNC-MAC achieves gains of 34% for an m_{low}-m_{high} UE pair ($K = 80$), while the gain reaches 38% for the m_{high}-m_{high} pair. Remarkably, the energy efficiency remains unaffected by the cell congestion levels, mainly due to the fact that i) longer idle intervals occur, when packet arrival rates decrease, and ii) the energy consumption in idle and reception state is similar.

3.5.3 Impact of LTE-A network deployment on ACNC-MAC performance

In this section, we study the effect of various LTE-A network parameters, i.e., MCSs and downlink packet scheduling policies, on ACNC-MAC performance, and the influence of idle UEs' distributions in a video transmission scenario.

3.5.3.1 Effect of MCS choice in downlink transmissions

Revisiting Figs. 3.7 and 3.8 in Section 3.5.2.3, we may observe that the performance of the ACNC-MAC protocol is affected by the MCSs utilized for the downlink transmission of the active UE pair.

More specifically, regarding the achieved throughput levels depicted in Fig. 3.7, we can see that the throughput of m_{high} UEs is significantly better than the throughput of m_{low}-m_{high} UEs. This observation can be explained by the fact that when higher order MCSs are used, the achieved downlink data rates are higher, leading to the increase of the packet arrival rates and creating more NC opportunities during the cooperation phase of the ACNC-MAC protocol.

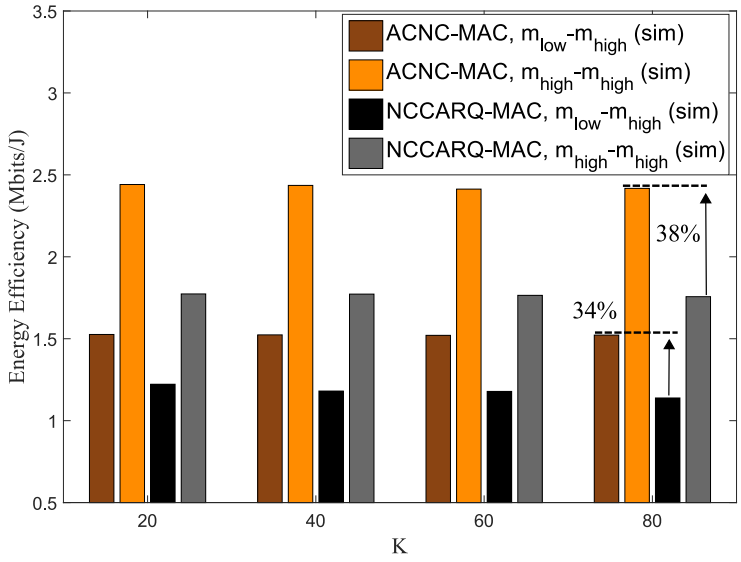


Figure 3.8: D2D energy efficiency for different SNR classes vs. K

Furthermore, in Fig. 3.8, we observe that the energy efficiency for the case of m_{high} UEs is higher than that of m_{low}-m_{high} UEs. More NC packets are transmitted when UEs with high packet arrival rates communicate using the ACNC-MAC protocol. When MCSs of lower order are used, the relays retransmit only one packet more often, thus more transmission rounds are required in order to deliver the same amount of data.

3.5.3.2 Effect of downlink packet scheduling policy

Aiming to investigate the influence of the utilized downlink packet scheduling policies on the D2D communication performance using the ACNC-MAC ptocol, we implemented three different scheduling policies, namely round robin (RR), maximum throughput (MT) and proportional fair (PF) [17]. In our study, the RR scheduler is used as the baseline. The MT scheduler maximizes the total throughput of the cell by prioritizing UEs with the best downlink channel SNRs. The PF scheduler aims to find a balance between overall throughput maximization and fairness by concurrently allowing all UEs to receive at least a minimal amount of RBs.

Inspecting Fig. 3.9, we see that the utilized scheduling policy affects the D2D throughput, although this influence differentiates according to the SNR class of the active UEs. Particularly, for the m_{high} UE pair, the MT scheduler achieves higher throughput than the other schedulers, even in high cell congestion, reaching an improvement of 12% ($K = 60$) and 190% ($K = 80$), comparing to PF and RR, respectively. In contrast, for the m_{low}-m_{high} UE pair, the PF scheduler improves the throughput, achieving an increase of 24% ($K = 20$) and 43% ($K = 40$), comparing to MT and RR schedulers, respectively.

Continuing, we can also observe that the MT scheduler is favorable for the m_{high} pair. Additionally, for the m_{low}-m_{high} pair, the throughput is higher using the PF scheduler.

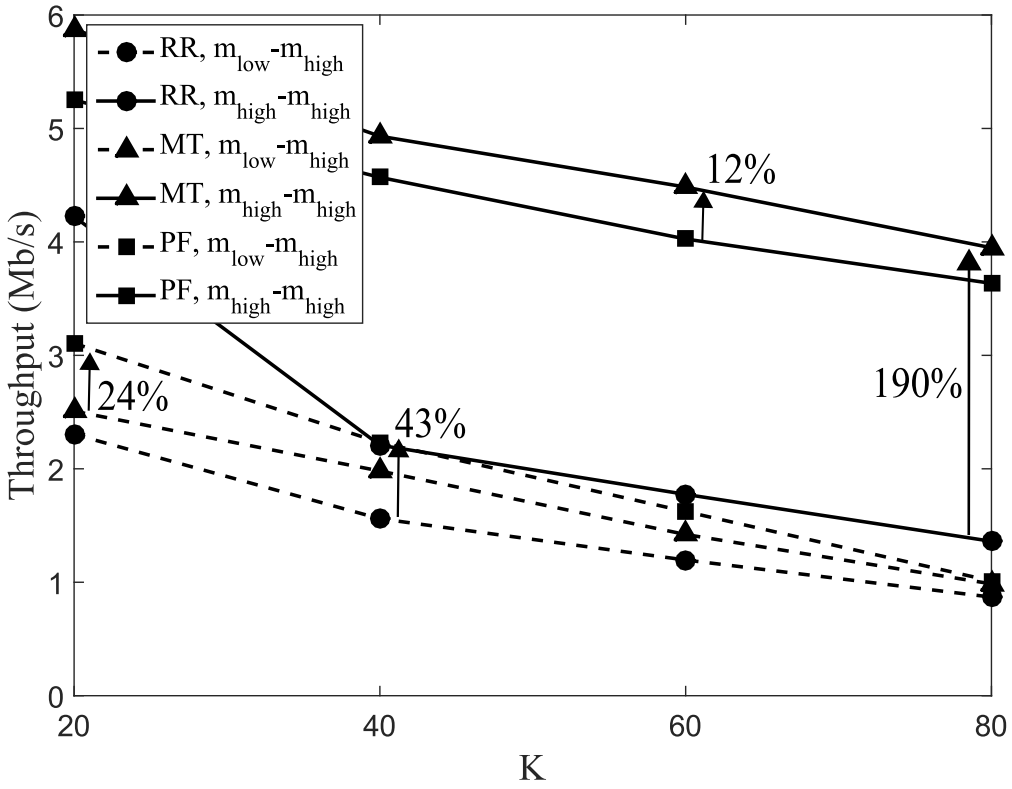
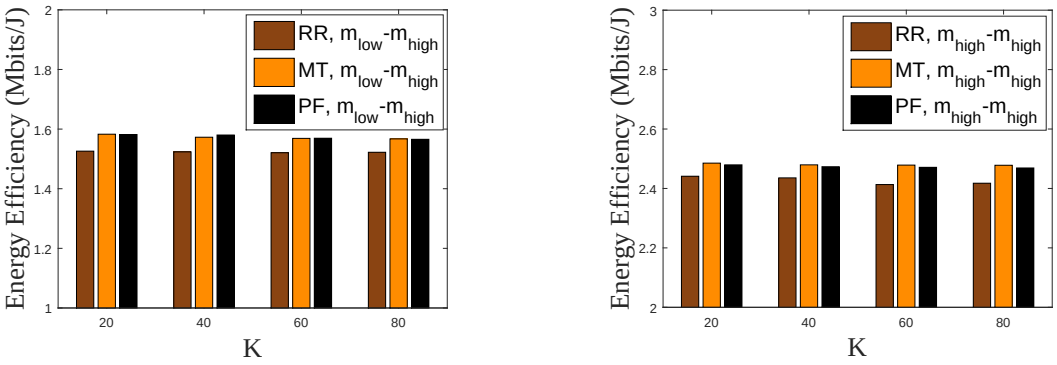


Figure 3.9: D2D throughput vs. K for different downlink packet scheduling policies



(a) Energy efficiency for m_{low}-m_{high} UE pair (b) Energy efficiency for m_{high}-m_{high} UE pair

Figure 3.10: D2D energy efficiency vs. K for different downlink packet scheduling policies

These observations are justified by the way RBs are allocated to UEs. More specifically, the MT scheduler allocates more RBs to the m_{high} UEs. The prioritization of these UEs in resource allocation induces higher packet arrival rates for them. In contrast, the PF scheduler treats the m_{low} UEs more fairly. It allocates to them a higher number of RBs than MT scheduler does, thus they experience higher packet arrival rates comparing to the other schedulers.

Unlike the D2D throughput, different trends are observed in the D2D energy efficiency behavior, as shown in Fig. 3.10. Remarkably, for both UE pairs under study, all schedulers result in similar energy efficiency. Actually, the increase of K reduces the packet arrival rates, inducing longer idle periods and more unfruitful communication rounds, as packet arrivals become quite scarce. Nevertheless, the similar energy consumption levels in idle and reception state lead to similar energy efficiency levels, independently of the scheduling policy and the cell congestion levels.

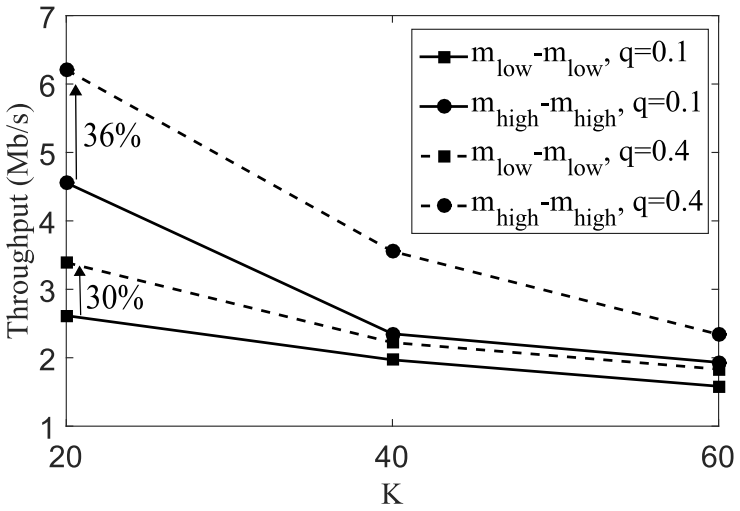
3.5.3.3 Effect of different idle UEs-relays proportions

In the previous scenarios, we have set a specific number of idle UEs that act as relays, performing the cooperative transmissions. In this section, we modify the proportion of the idle UEs (relays). More specifically, we evaluate the ACNC-MAC protocol using numbers of relays equal to 10% and 40% of $K \in \{20, 40, 60\}$ and defining their proportion as $q \in \{0.1, 0.4\}$.

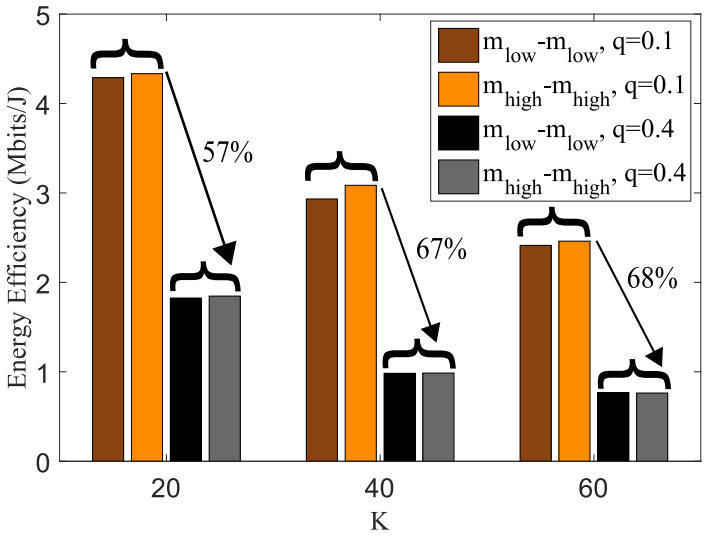
As expected, the achieved throughput demonstrates a downward trend as K increases, independently of the MCS used (Fig. 3.11(a)). Nevertheless, the throughput of m_{high} UEs for each K is higher than the throughput of m_{low} UEs, which are disfavored even when the number of relays increases. In any case though, the throughput performance of the ACNC-MAC protocol seems to improve when more relays exist, e.g., comparing the cases of an m_{high} UE pair and an m_{low} UE pair ($K = 20$), the throughput is 36% and 30% higher, respectively, when $q = 0.4$. This effect can be attributed to the coexistence of fewer active UEs, which induces higher data rates, and the utilization of higher number of relays during the ACNC-MAC cooperation phase.

In Fig. 3.11(b), we observe that the energy efficiency reduces, when the cell becomes more congested. This is due to the fact that when more UEs are active, more time is required to deliver the video sequence. Still, the energy efficiency performance for both UE classes is better with $q = 0.1$, i.e., when fewer relays participate in the cooperation phase. In case that a higher number of relays are used, the total energy consumption of the D2D network increases. Therefore, the energy efficiency is significantly lower, when $q = 0.4$, whereas in average, the decrease of energy efficiency reaches 57%, 67% and 68% for $K = \{20, 40, 60\}$, respectively. It should be also noted that the increased throughput of the scenarios with $q = 0.4$ does not improve energy efficiency due to the high energy consumption of the relays.

Additional information about the impact of relays distribution on energy consumption can be derived by inspecting the battery drain levels of the active UEs, depicted in

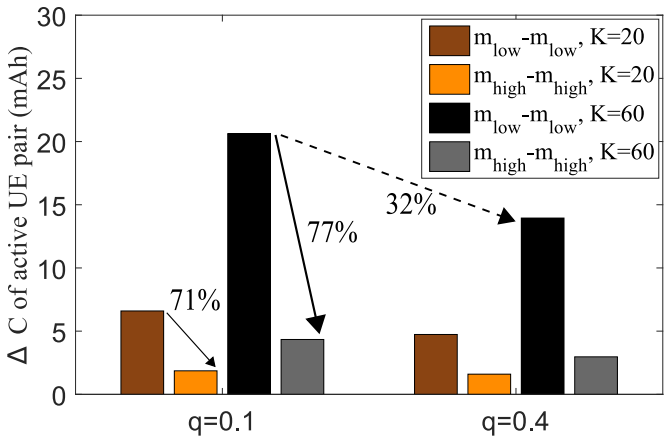


(a) D2D throughput vs. K

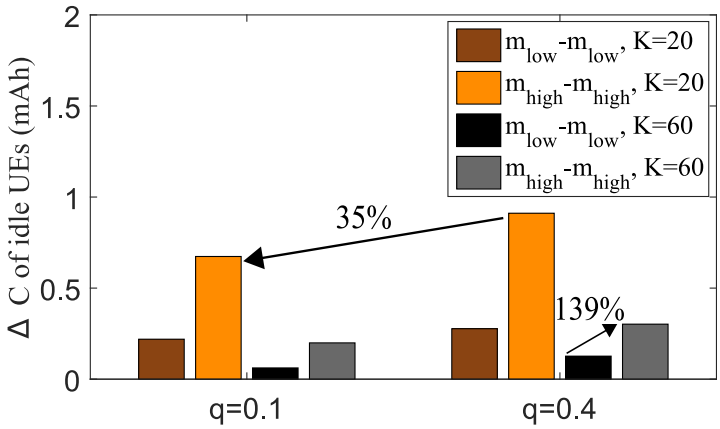


(b) D2D energy efficiency vs. K

Figure 3.11: Impact of different idle UEs proportions on ACNC-MAC throughput and energy efficiency



(a) Average battery drain of active UE pair



(b) Average battery drain of idle UEs

Figure 3.12: Impact of different idle UEs proportions on ΔC of UEs using ACNC-MAC

Fig. 3.12(a). We may see that the UE pair's ΔC is higher when lower order MCS is used, as the downlink video transmission lasts longer due to lower data rates. For instance, when $q = 0.1$, the ΔC of an m_{high} UE pair is 71% ($K = 20$) and 77% ($K = 60$) lower than that of an m_{low} UE pair, respectively. Equally perceptible are the differences between the two idle UEs proportions with regard to the energy consumption of the active UE pair. Considering the case of an m_{low} UE pair ($K = 60$), the increase of q from 0.1 to 0.4 causes a diminution of 32% of the ΔC of the active UE pair. A possible interpretation of this result is that the benefit from the shorter transmission duration when fewer active UEs exist is outweighed by the D2D communication overhead.

Focusing on the ΔC levels of the relays, illustrated in Fig. 3.12(b), we can see some different trends from those observed in the ΔC levels of the active UE pair. The battery of the relays reduces to a greater extent if the transmissions of m_{high} UEs are served, e.g., for $K = 60$ and $q = 0.4$, the relays' ΔC is 139% higher than the ΔC when an m_{low} UE pair exchanges data. It seems that the throughput improvement of m_{high} class is accompanied by an increase in the energy consumption of the relays, as the frequency of packet arrivals is higher and packet retransmissions occur more frequently. Moreover, the ΔC of the relays is higher when more idle UEs are used, as more relays contend for channel access during the cooperation phase. For instance, in case of an m_{high} UE pair ($K = 20$), the increase of the q value, i.e, the proportion of idle UEs used as relays, leads to 35% higher ΔC for the relays.

3.6 Chapter concluding remarks

In this chapter, a cooperative NC-based MAC protocol (ACNC-MAC) for outband D2D bidirectional communication in LTE-A cell has been introduced. An analytical model for ACNC-MAC throughput performance in saturated network conditions has been presented, along with the throughput analytical model that incorporates characteristics of both LTE-A and D2D links. We have assessed the ACNC-MAC performance in saturated and non-saturated network conditions and also, in the heterogeneous cellular-D2D system under different network setups.

The conducted simulations have revealed that the ACNC-MAC protocol is beneficial in terms of both throughput and energy efficiency comparing to the SoA. More specifically, when the D2D network operates in saturated conditions, ACNC-MAC offers up to 73% higher throughput (PER equal to 0.5) and 71% higher energy efficiency (PER equal to 0.3). In case of Poisson packet arrivals (non-saturated conditions), the improvement of throughput and energy efficiency reaches up to 41% and 39%, respectively (PER equal to 0). Additionally, when the D2D pairs experience high downlink data rates, the throughput achieved by ACNC-MAC is up to 226% higher, whereas the energy efficiency is up to 38% higher, comparing to the SoA.

It has been also observed that the D2D throughput improves when more relays are used, reaching an increase of 36% when 8 relays are used instead of 2 (assuming 20 active

UEs in the cell). However, in the same scenario, the energy efficiency reduces by 57%. Considering this result, we should mention that although the use of more relays improves the D2D throughput, their number should be properly selected in order to avoid excessive battery consumption that would decrease the energy efficiency. This effect may hinder the willingness for cooperation of the idle UEs that can be used as relays for the D2D communication of a UE pair.

Furthermore, regarding the coexistence of cellular and outband D2D communication links, our study has shed some light on cellular network-related factors that affect the outband D2D performance and the tradeoffs that arise. More specifically, it has been shown that the effect of scheduling policies varies with the cellular channel quality of the active UEs. Consequently, each scheduling policy is suitable in different cases, i.e., UEs with high downlink SNRs experience higher throughput with the MT scheduler (up to 190% increase comparing to RR scheduler), whereas for UEs with poor downlink channel conditions, e.g., in urban environments with obstacles, the PF scheduler is preferable (up to 43% increase comparing to RR scheduler). As a final remark, we should note that the benefit of using MCSs of higher order in cellular transmissions is depicted on the D2D performance, even when the cell congestion increases.

3.A Appendix

3.A.1 Proof of lemma 1

As described in Section 3.3, in the proposed ACNC-MAC protocol, a packet arrival to at least one of the two UEs consisting the D2D pair under study initiates a new transmission round. Consequently, in a random slot, either of the following events occur: i) no packet arrives at the queue of any UE, ii) a packet arrives at the queue of either of the two UEs, and iii) packets arrive at the queues of both UEs. Thus, a new transmission round will begin when either of the events ii) and iii) occurs. Assuming that packets arrive at a UE z according to Poisson distribution with rate λ_z, the probability that one or more packets arrive in a time slot is given by [105]:

$$P_z = 1 - \mathrm{e}^{-\lambda_z \mathbb{E}[T_{slot}]}. \tag{3.36}$$

When the event D_2 occurs, packets arrive at both UEs, thus the probability of packet arrivals $P(D_2)$ is given by the multiplication rule, as the product of $(1 - \mathrm{e}^{-\lambda_1 \mathbb{E}[T_{slot}]})$ and $(1 - \mathrm{e}^{-\lambda_2 \mathbb{E}[T_{slot}]})$, which are the probabilities of packet arrival in UE_1 and UE_2, respectively.

Additonally, the term $\mathbb{E}[T_{slot}]$ can be mathematically expressed as [105]:

$$\mathbb{E}[T_{slot}] = (1 - p_{tr})\sigma + 2p_s T_s + p_c T_c, \tag{3.37}$$

where σ is the idle slot duration, while $T_s = \mathrm{DIFS} + T_{pkt}$ is the duration of transmission of a packet by a UE and $T_c = \mathrm{DIFS} + T_{pkt} + \mathrm{SIFS} + T_{RFC}$ is the expected time of collision. The probability that an active UE successfully transmits is $p_s = 2\tau(1 - \tau)$, where τ is

Table 3.2: Values of x and y terms of Eq. (3.32)

Case *(i,j)*	$x_{i,j}$	$y_{i,j}$
(0,1)	SIFS+T_{ETC}	0
(0,2)	SIFS+T_{ETC}	T_{pkt}
(1,1)	SIFS+T_{ETC}	T_{ACK}+SIFS
(1,2)	SIFS+T_{ETC}+T_{pkt}	T_{pkt}+T_{ACK}+SIFS
(2,2)	SIFS+T_{ETC}+T_{pkt}^{NC}	T_{pkt}+2(T_{ACK}+SIFS)

the probability that a UE attempts to transmit in a random slot. The probability that at least one of UE_1 and UE_2 transmits is p_{tr}, whereas the two UEs experience a collision with probability $p_c = \tau^2$. The probabilities p_{tr}, p_s and p_c are calculated by solving the system of τ and the Markov chain's stationary probability at the initial state, $b_{0,0}$. In our case, the probability of having at least one packet to any of the two active UEs utilized by τ and $b_{0,0}$ can be derived as in [105] by setting $\lambda = (\lambda_1 + \lambda_2)/2$.

3.A.2 Proof of lemma 4

At each communication round, $|M_i|$ out of N relay candidates contend for channel access and their transmissions may result in collision. The expected value of $|M_i|$ for each case i expresses the number of relays that have received i packets. The probability $P(H_i)$ of each ACNC-MAC case is:

$$P(H_i) = \sum_{j=1}^{2} P(H_{i,j}), i \in \{0, 1, 2\}. \tag{3.38}$$

Similarly as in Section 3.4.2, we derive the probabilities $P(H_i) \forall i$ as follows:

1. *Case 0*: No relay has received any packet, thus all relays belong to M_0 ($N = k$). Using Eq. (3.22), $P(H_0)$ is given by:

$$P(H_0) = P(H_{0,1}) + P(H_{0,2}). \tag{3.39}$$

2. *Case 1*: Relays with either one packet or zero packets exist. Even if packets from both UEs are transmitted, none of the idle UEs has correctly received both of them. Hence, $|M_1|$ is equal to the number of relays that have one packet. Using Eq. (3.24), $P(H_1)$ is given by:

$$P(H_1) = P(H_{1,1}) + P(H_{1,2}). \tag{3.40}$$

3. *Case 2*: The active relay set M_2 contains the relays that have received both packets and can perform NC. From Eq. (3.28), $P(H_2)$ is given by:

$$P(H_2) = P(H_{2,2}). \tag{3.41}$$

Substituting Eqs. (3.39)-(3.41) in Eq. (3.38) yields the values $P(H_i), i \in \{0, 1, 2\}$, which are required for the estimation of $P(|M_i| = k)$.

3.A.3 Proof of lemma 5

In the first component, i.e., $\mathbb{E}\left[T_{i,j}^{min}\right]$, the term $\mathbb{E}\left[T_{init}\right]$ is the delay induced by the initial contention phase between the active UEs. The retransmission duration $x_{i,j}$, in case that the relays are perfectly scheduled and collisions do not occur, varies according to the number of retransmitted packets. Similarly, the additional time $y_{i,j}$ consumed in a contention-free cooperation phase differentiates according to the number of delivered packets, representing the number of ACK frames expected. The values $x_{i,j}$ and $y_{i,j}$ are reported in Table 3.2.

The second component, i.e., $\mathbb{E}\left[T_i^{cont}\right]$, is the delay caused by the relays' contention, expressed as the product of $\mathbb{E}\left[r\right]$ and $\mathbb{E}\left[Tc_i\right]$. $\mathbb{E}\left[r\right]$ is the expected number of retransmissions required for the successful reception of all packets by their destinations and is estimated as a function of $PER_{(UE_1 \leftrightarrow r)}$ and $PER_{(UE_2 \leftrightarrow r)}$ [103]. $\mathbb{E}\left[Tc_i\right]$ is the expected time needed for packets transmissions during the relays' contention.

For the calculation of the $\mathbb{E}\left[Tc_i\right]$ values, the backoff counter model in [40] is applied. As already explained, the relays select their backoff times from different ranges that are dictated by the number of overheard packets. Hence, different values of the average time until a relay transmits successfully must be considered in correspondence with the ACNC-MAC cases. To that end, the value of $\mathbb{E}\left[Tc_i\right]$ $\forall i$ can be estimated as:

$$\mathbb{E}\left[Tc_i\right] = \left(\frac{1}{p_i^{suc}} - 1\right) \cdot \left[\left(\frac{p_i^{idle}}{1 - p_i^{suc}}\right)\sigma + \left(\frac{p_i^{col}}{1 - p_i^{suc}}\right)T_i^{col}\right], \tag{3.42}$$

where p_i^{suc}, p_i^{idle}, p_i^{col} are the probabilities of having a successful, idle or collided slot [40]. The probabilities utilized by the Bianchi model must be computed separately for each ACNC-MAC case C_0, C_1 and C_2 using the respective active relay set size estimations $|M_0|$, $|M_1|$ and $|M_2|$, which are derived in Section 3.4.2.3. Identically, the duration of the collision among the transmissions of different relays T_i^{col} is different $\forall i$ and is given by:

$$T_0^{col} = \text{SIFS} + T_{ETC}, \tag{3.43}$$

$$T_1^{col} = \text{SIFS} + T_{ETC} + T_{pkt}, \tag{3.44}$$

$$T_2^{col} = \text{SIFS} + T_{ETC} + T_{pkt}^{NC}. \tag{3.45}$$

Chapter 4

The SCD2D–MAC protocol for integration of social awareness in outband D2D communication

4.1 Introduction

Following the proliferation of social networks and cutting-edge mobile devices, social ties among users can promote D2D cooperation. In D2D cooperative communication, multiple devices in close proximity attempt to access the wireless medium. As already discussed, their interactions at medium access level are affected by the social features of the mobile users, as socially connected users are more likely to engage in D2D cooperation. Moreover, the energy consumption of power–constrained mobile devices affects the effectiveness of D2D cooperative communication, stressing the need for incorporating energy awareness in D2D networking.

Taking into account the aforementioned context and the characteristics of modern social networking scenarios, in this chapter, we investigate the implications of green D2D cooperation from a social-aware perspective and provide intuition towards their resolution. To this end, the contribution of this chapter can be summarized in the following points:

(i) We focus on the integration of social awareness in green D2D-MAC design. More specifically, we present a social-aware cooperative D2D MAC protocol (SCD2D-MAC) that promotes cooperation among socially related neighboring users and evaluate it in D2D networking scenarios. SCD2D-MAC exploits social awareness in order to improve the energy efficiency of D2D cooperative communication. The

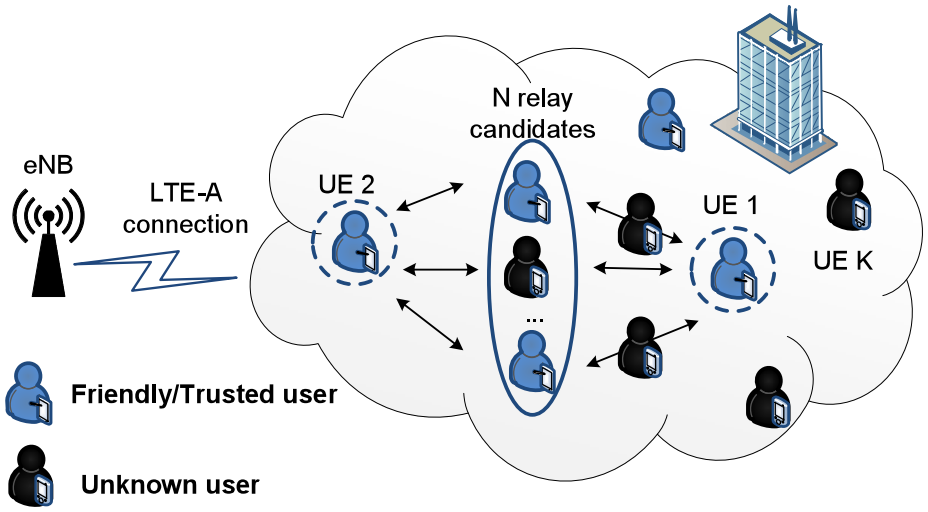


Figure 4.1: D2D enabled LTE–A network

performance assessment of the proposed protocol reveals that significant gains can be achieved in terms of energy consumption without hindering the content exchange completion time, when social features are considered.

(ii) We outline the practical concerns that arise, from the network and the users' perspective, by the adoption of social awareness in green D2D cooperation. The discussed issues may hinder the actual benefits of social-aware design of green D2D cooperation and should be taken into account when D2D cooperative structures are orchestrated.

The remainder of the chapter is structured as follows. In Section 4.2, the considered system model is described. In Section 4.3, the SCD2D-MAC is presented in detail, whereas its performance is evaluated in Section 4.4. In Section 4.5, several practical issues that arise in the integration of social awareness in D2D cooperation are discussed. Last, Section 4.6 provides some concluding remarks on this chapter.

4.2 System model

In the LTE-A network depicted in Fig. 4.1, the UE pair that consists of UE_1 and UE_2 intends to initiate a bidirectional communication among them in order to exchange data, which either are concurrently downloaded via cellular links that the UEs can maintain or may already exist in the UEs before the initiation of the D2D exchange. In the considered cell network, a total number of W RBs is available and a total number of K active UEs reside in the cell.

Regarding the cellular connections, we assume that the UEs are located in various distances from the eNB. A fixed transmission power $P_{\text{trans}}^{\text{eNB}}$ from the eNB is used, whereas the RBs that serve the downlink transmissions are allocated to the UEs according to the round robin scheduling policy. Hence, the UEs may experience different SNR levels, which

determine the MCSs preferred for the downlink transmissions. For the proper selection of MCS utilized in cellular transmission, the experienced SNR of each eNB-UE link must be estimated. Letting N_0 be the noise power spectral density in dBm/Hz for each link and $P_{\text{trans}}^{\text{eNB}}$ the transmission power of the eNB, then the SNR is given as:

$$SNR[dB] = P_{\text{trans}}^{\text{eNB}} - P_l - N_0, \tag{4.1}$$

where the path loss component P_l can be computed using the modified COST231 Hata urban propagation model [113], considering an urban macro environment, as follows:

$$P_l = 34.5 + 35 \log_{10}(d), \tag{4.2}$$

where d is the distance between a UE and the eNB.

During the D2D data exchange, in the considered network, erroneous packet transmissions might occur due to the fluctuations of the quality of the D2D links. If a UE fails to decode a packet, it may ask for cooperation from UEs in close proximity, which are able to opportunistically overhear the packets exchanged during the $UE_1 \leftrightarrow UE_2$ communication. As the UEs maintain social ties with other UEs via their social networking applications, they prefer to utilize friendly UEs as relays. Out of the K UEs that exist in the cell a number of N UEs are considered to be relay candidates and may either be socially connected with the UE pair (friendly relays that exist in the pair's social preference list) or be totally unknown to the UE pair.

Regarding the channel model, it is assumed that the wireless channels between the UEs and their relays are assumed to be independent of each other. We denote as PER the packet error rate that characterizes each of the D2D links between the UEs and the relays. In the Wi–Fi interface, we denote as P_{Rx}, P_{idle} and P_{Tx} the power level for the reception, idle and transmission mode, whereas the UEs and the relays exchange data using a transmission data rate R_{Tx}.

4.3 The SCD2D-MAC protocol design

An overall inspection of the challenges of the considered social networking scenarios shows that social awareness can improve D2D networking and make the traits of cooperating over unlicensed spectrum more appealing to the users. Taking into account the context of social networking and the energy consumption issues that arise by the social awareness, we present a social-aware cooperative D2D MAC (SCD2D-MAC) protocol as a paradigm of incorporation of social information in D2D MAC design that can improve the D2D energy efficiency. SCD2D-MAC promotes cooperation among users with social ties in case of D2D communication between a pair of users, reducing the overall energy consumption of D2D cooperative communication.

The main functionality of SCD2D-MAC relies on the availability of social context information to the UE pair, i.e., the UE pair should be aware of the friendly UEs that reside in close proximity and can be used as relays. For this purpose, the eNB determines

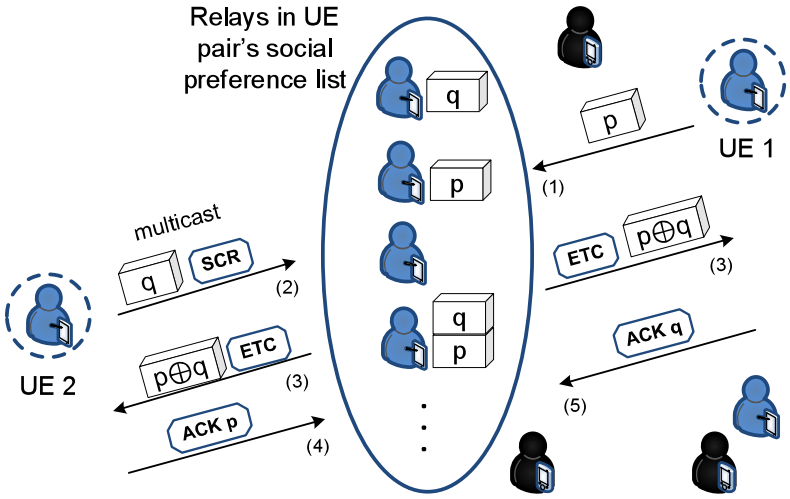


Figure 4.2: D2D cooperative data exchange with SCD2D-MAC

D2D candidates during the peer discovery phase, by paging possible friendly users and determining D2D pairs. After D2D connections are established, a social preference list for each D2D pair is constructed, including "identification details" of users eligible for D2D communication with each pair, and is sent to the pair by the eNB. Friendly users opportunistically encountered in vicinity can serve as relays. Once the peer discovery phase and the initialization of D2D connections are completed, the UE pair exchange data using the SCD2D-MAC protocol.

The existence of social connection between the relays and the pair is the criterion that determines the decision of relays to engage in D2D cooperation. We assume that neighbouring UEs are either friends of a UE pair or unknown users. Users belonging to the pair's social network or are members of the same online community are willing to help the pair's communication. Conversely, unknown relays are not bound to cooperate and are likely to content for channel access aiming to serve D2D transmissions for their own benefit. Thus, their participation in D2D cooperation might cause a series of unfruitful communication rounds, from the pair's viewpoint. As more D2D transmissions might be required to deliver the pair's data due to unknown relays' intervention, the energy efficiency of D2D cooperation might deteriorate.

Let us consider the D2D pair of UE 1 and UE 2, who desire to exchange data directly using Wi-Fi, after having obtained the social preference list (Fig. 4.2). At each communication round, a user gains channel access using the DCF of the IEEE 802.11 standard specification [27] and transmits its packet (step 1). The other user fails to decode it correctly and in step 2, it sends a *social-cooperation-request* (SCR) packet to request for cooperation from adjacent users-friends. In the D2D pair's Wi-Fi range, two types of users may co-exist:

(i) Users without social ties with the pair.

(ii) Users that maintain social connections with the pair.

Preferably, users-contacts of the pair are utilized as relays. The SCR packet contains the necessary information for the identification of the pair by the possible relays. It should be noted that in social-unaware D2D MAC protocols, the contingency that friendly and unknown users coexist is not explicitly handled. Thus, the relay candidates gain channel access equitably, according to the IEEE 802.11 rules, regardless of the users' social ties. In SCD2D-MAC, the social dimension of D2D cooperation is reflected in relay selection process, which prioritizes the use of friendly users as relays. More specifically, in the cooperation phase, the SCR packet is transmitted to a multicast group that consists exclusively of users-friends of the pair. Only the users in this group are considered to be trusted and receive the SCR packet, which indicates their eligibility as relays.

After distinguishing the friendly relays and organizing them in a multicast group, the SCD2D-MAC protocol prioritizes them according to the number of packets they manage to decode, improving the D2D cooperation performance. In each cooperation round, a relay may overhear up to two packets, namely it may receive either packets from both users in the D2D pair and can perform NC, or only one packet (the packet of one user, either UE 1 or UE 2) or it may not be able to correctly decode any packet. Each relay that wishes to transmit uses a backoff counter, as required by the DCF method. The relay prioritization is accomplished using non-overlapping ranges for the backoff counter of the relays. The backoff range is divided into several ranges according to the number of packets existing in each relay, in a way that relays with more packets can gain channel access. For instance, relays with both packets select their backoff counter from a backoff range with lower values than those used by relays with one packet.

With regard to the number of packets received by the eligible relays, the cooperation phase may lead to one out of three possible outcomes. First, if at least one of the relays receives packets of both users, namely packets p and q, network coding can be performed. In this case, an encoded packet is transmitted by the relay (step 3 in Fig. 4.2). Second, if no relay receives both packets but there exist relays with one packet, either p or q, the selected relay transmits the packet it has received. Last, there is the contingency that no packets are correctly decoded by any friendly adjacent user, leading to an unfruitful cooperation round. Subsequently, the selected relay indicates the number of packets it will transmit in the *eager-to-cooperate* (ETC) packet, sent along with data packets correctly decoded. Once ETC is received, the pair is aware of the number of ACKs that will terminate the cooperation phase. For the example of Fig. 4.2, two ACKs are transmitted, indicating the successful reception of p and q (steps 4 and 5). If no data packet is transmitted, the cooperation ends with the ETC transmission.

It should be also noted that SCD2D-MAC does not require a metric that quantifies that strength of the social ties between the D2D pair and the relay candidates. The capability of a friendly relay to perform NC during the cooperation phase is the factor that finally determines the selection of a relay. For instance, considering the case where a relay candidate r_1 has loose social ties with a D2D pair and is closer to UE 1 and a relay candidate r_2 has strong social ties with the pair but is closer to it, priority will be

given to the transmission of the relay candidate that is able to perform NC. Both UEs r_1 and r_2 are considered to be "equally friendly" to the D2D pair. In this case, the protocol would choose the relay that is able to perform NC. If neither of the two UEs were able to perform NC, the protocol would choose either of them as relay.

4.4 Performance assessment

We quantitatively evaluate the SCD2D-MAC protocol under the influence of information about users' social structures in the D2D cooperative communication scenarios of a socially connected pair of users that exchange data of user or cellular network origination. Aiming to highlight the effect of social characteristics in D2D cooperation, we compare SCD2D-MAC with two SoA protocols that do not consider the social dimension, i.e., the ACNC-MAC protocol [114] and the NCCARQ-MAC protocol [40], considering different proportions of friendly relays within the pair's range.

We have developed a C++ simulator that implements the three protocols. The D2D cooperative communication performance is assessed in terms of data exchange completion time, namely the time required for successful reception of exchanged content by both users. Furthermore, we estimate the energy efficiency [111] and the average battery drain [112] of the D2D network, considering the energy consumption of all participating users.

4.4.1 Simulation setup

The D2D pair in Fig. 4.2 resides in the coverage area of an LTE-A cell with $K = 30$ active users, out of which a number of $N = 20$ users are relay candidates. They either maintain social ties with the pair or are strangers. We define as $\alpha \in \{0.2, 0.4, 0.7, 0.9\}$ the proportion of friendly relays in the pair's area, corresponding to 20%, 40%, 70% and 90% of the relay candidates' number.

As already discussed, SCD2D-MAC distinguishes the friendly relays by explicitly asking for their cooperation. Conversely, the ACNC-MAC and NCCARQ-MAC protocols cannot perform relay discrimination, allowing the use of any adjacent user as relay. Hence, there exists the risk that unknown relays may gain channel access and serve transmissions of their own interest. With NCCARQ–MAC, the cooperation phase begins only if the relays receive packets from both users and can perform NC, whereas with ACNC-MAC, cooperation may be initiated even with fewer packets at the relays.

All protocols are tested in two D2D communication scenarios, denoted as A and B, using the settings in Table 5.2. The users' devices are equipped with batteries of initial capacity equal to 1300 mAh and LTE-A and Wi-Fi radio interfaces that can be used simultaneously. In the presented results, a fixed PER is used for all D2D links, as different PER values influence the protocols' performance as anticipated, without affecting our conclusions. In scenario A, the two users exchange two files of 5 MB size, already existing in their devices and the network operates under saturated conditions. In scenario B,

Table 4.1: Simulation parameters for performance evaluation of SCD2D-MAC protocol

Cellular network parameters (scenario A)	
Parameter	**Value**
W	100 RBs (20 MHz)
Resources scheduling	Round robin
N_0 (dBm/Hz)	-174 dBm/Hz
$P_{\text{trans}}^{\text{eNB}}$	46 dBm
UEs-eNB distance d	700-800 m
Modulation scheme	64-QAM
TTI	1 ms
P_{Rx}	2 W
Video sequence	Foreman, QCIF, 15 fps
D2D network parameters (both scenarios)	
Parameter	**Value**
MAC+PHY header	52 bytes
Time slot	10 μs
R_{Rx}	54 Mb/s
SCR	16 bytes
Packet payload size	512 bytes
ETC, ACK	14 bytes
PER	0.2
$P_{Rx} = P_{idle}$	1.34 W
P_{Tx}	1.9 W

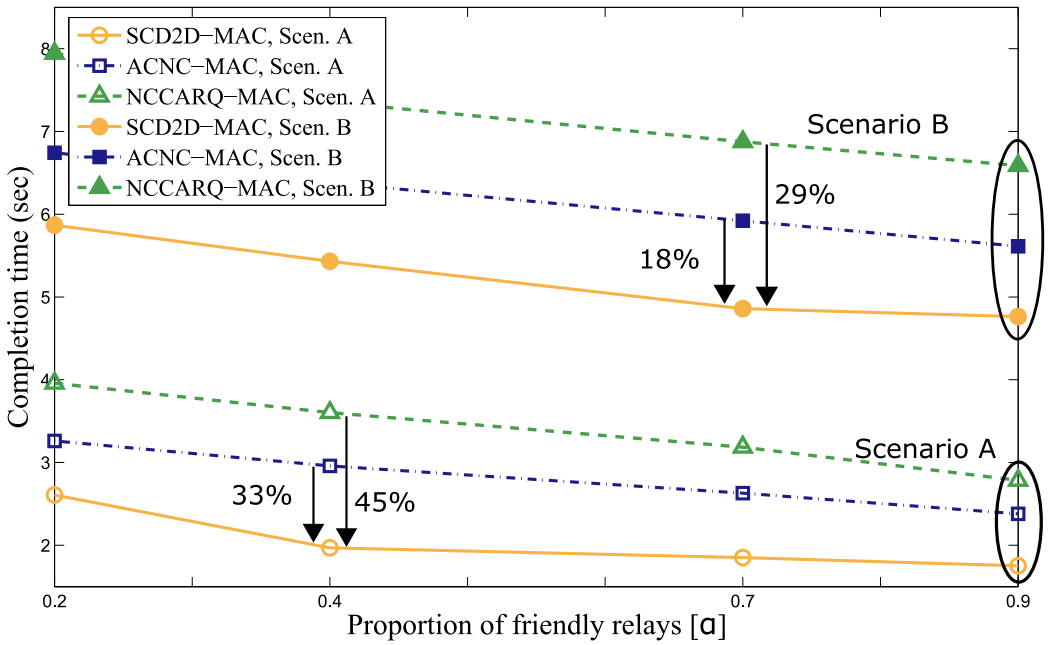


Figure 4.3: D2D content exchange completion time

the users exchange video content they receive from cellular connections. Therefore, the resources scheduling policy for downlink transmissions determines the packet arrival rate at the UE pair, creating non-saturated conditions.

4.4.2 Performance results

In Fig. 4.3, the data exchange completion time achieved by the three protocols is depicted. It can be clearly seen that the increase of the portion of friendly relays (α) improves the performance of all protocols, since fewer cooperation rounds are exploited by unknown users. However, both ACNC-MAC and NCCARQ-MAC need significantly higher time to complete the exchange than the SCD2D-MAC protocol. Indicatively, for $\alpha = 0.4$ in scenario A, SCD2D-MAC achieves 33% and 45% lower completion time than ACNC-MAC and NCCARQ-MAC, respectively. Similarly, in scenario B, the decrease of completion time with SCD2D-MAC reaches 18% and 29%, for $\alpha = 0.7$. This differentiation can be explained by the fact that SCD2D-MAC restricts the set of relays, explicitly asking for the cooperation of friendly users only. Thus, each cooperation round serves exclusively the pair's D2D transmissions.

The influence of α level in D2D cooperation performance is also perceptible in Fig. 4.4, which depicts the energy efficiency levels achieved by the three protocols under comparison in both D2D content exchange scenarios. We observe that as the α value increases, the energy efficiency reduces, since more relays are engaged in D2D cooperation. Hence, the total energy consumption in the D2D network increases. Due to this effect, even though the existence of more relays reduces the data exchange completion time in all cases, the

energy efficiency does not follow the same trend. However, it should be noted that the multicast functionality of SCD2D-MAC enables the use of friendly relays only, improving the energy efficiency, comparing to ACNC-MAC and NCCARQ-MAC. For instance, in scenario A, for $\alpha = 0.2$, the energy efficiency of SCD2D-MAC is 18% and 35% higher than that achieved by ACNC-MAC and NCCARQ-MAC, respectively (Fig. 4.4(a)). In scenario B ($\alpha = 0.4$), the resulting improvement reaches 10% and 17%, as shown in Fig. 4.4(b).

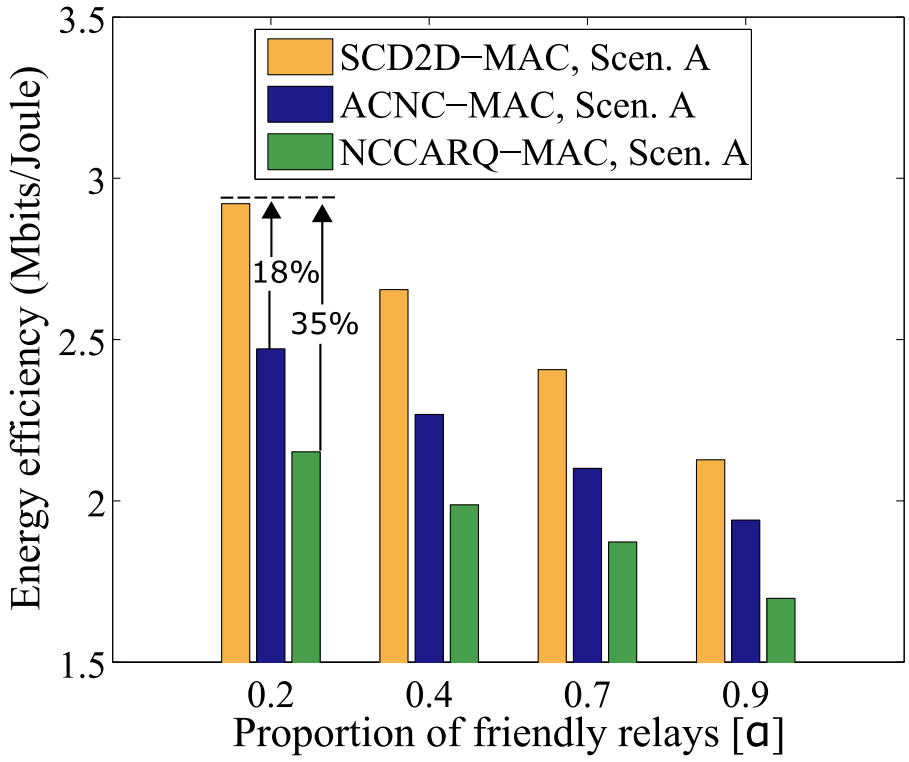
The D2D energy efficiency performance is in accordance with the battery usage levels illustrated in Fig. 4.5. More specifically, the average battery drain for the pair and the relays increases alongside with α, as a higher number of friendly relays contend for channel access in order to support the pair's communication. However, the use of SCD2D-MAC results in lower total energy consumption, as only a portion of neighboring users are selected to act as relays, transmitting data packets that are useful to the pair and reducing the completion time. Particularly, for $\alpha = 0.2$, the battery drain with SCD2D-MAC is 44% and 58% lower than ACNC-MAC and NCCARQ-MAC in scenario A and 29% and 37% lower in scenario B, as depicted in Fig. 4.5(a) and Fig. 4.5(b), respectively.

4.5 Practical issues in integration of social awareness in D2D cooperation

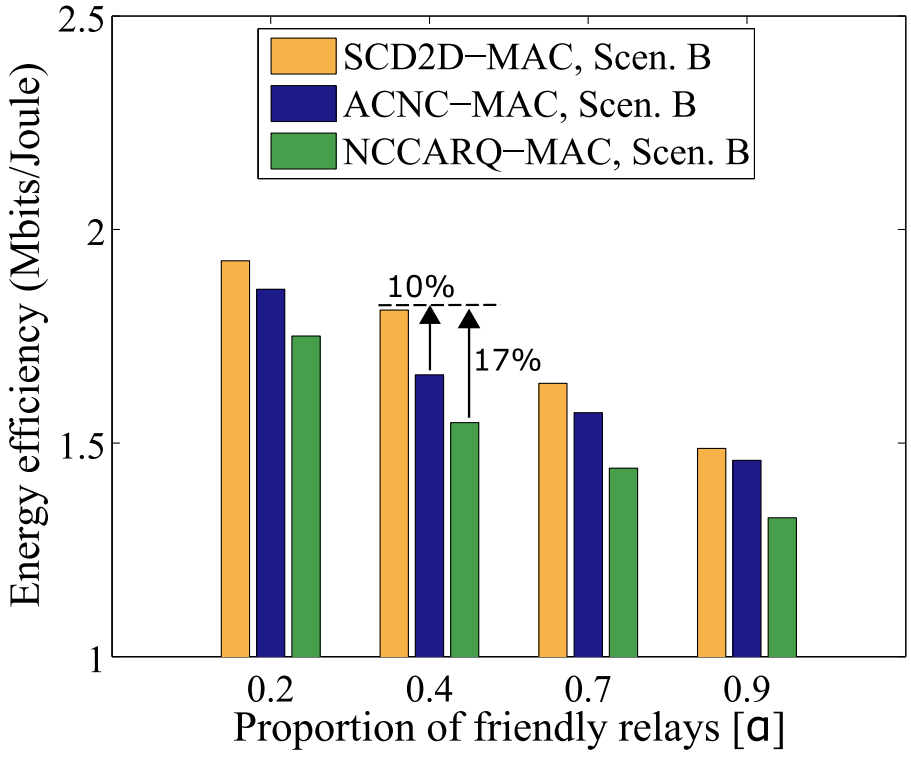
Promoting D2D cooperation among users with social ties is beneficial for the users' experience, in terms of data exchange completion time and battery drain. However, when the knowledge of social parameters is introduced in actual D2D cooperative networks, practical issues arise that may hinder opportunities for D2D cooperation and impact on D2D performance.

In realistic social-aware cooperative D2D networks, information of social domain about a possibly large number of users, e.g., in D2D data dissemination scenarios, is usually required, in order to obtain the users' social structures. This information can be transmitted to cellular infrastructure by the users' devices. During this process, additional signaling overhead in the cellular network elements that coordinate the D2D users is created. Without cellular network intervention, neighboring devices might have to exchange users' social information in an ad hoc manner, increasing the congestion in the D2D network. In any case, the benefits of social awareness in D2D cooperation should be studied in conjunction with the impact of additional network load that the transmission of users' social information induces.

To further harness the traits of using the knowledge of users' social ties in D2D cooperative structures, the network operators should provide practical incentives that can stimulate their mobile customers' interest in cooperation. However, motivating the users' participation is not trivial, as it requires observation of social characteristics and behavior in order to make the "remuneration" for D2D cooperation attractive. Although there exist approaches that integrate incentive mechanisms in D2D design, such as [31], it is usually assumed that all users are interested in the same type of payoff, e.g., monetary

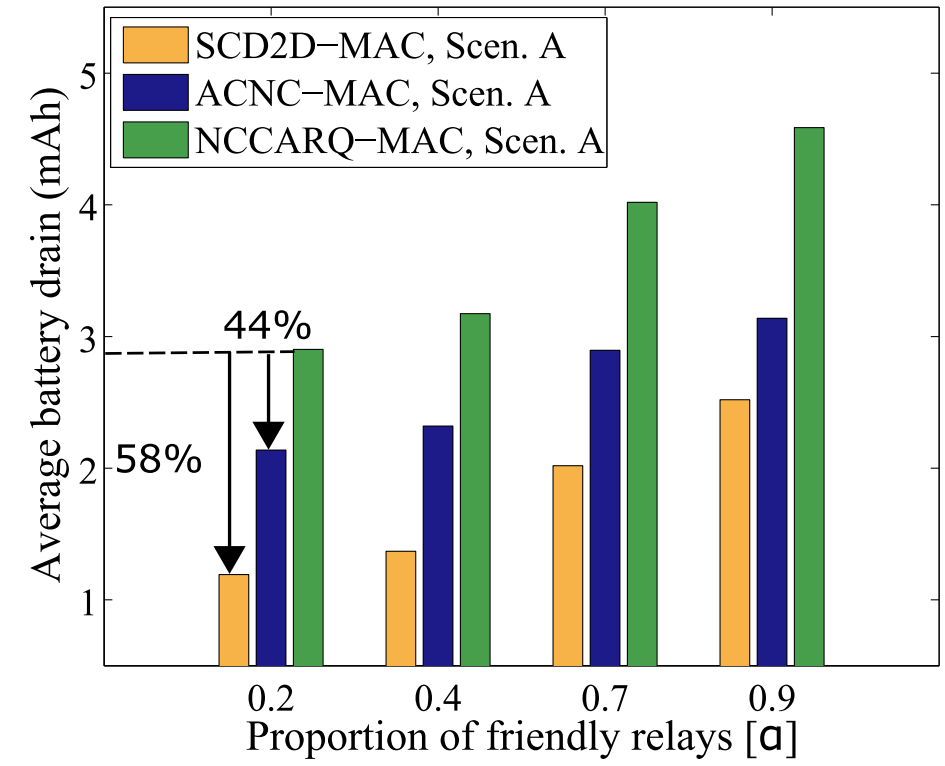


(a) Scenario A

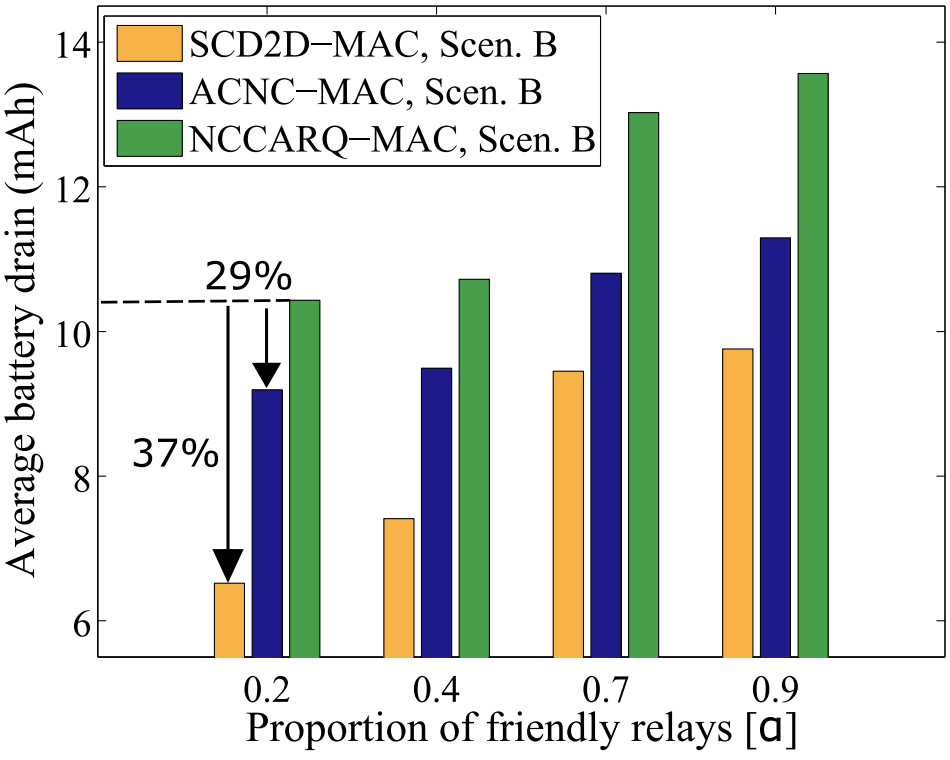


(b) Scenario B

Figure 4.4: Energy efficiency in D2D content exchange



(a) Scenario A



(b) Scenario B

Figure 4.5: Average battery drain in D2D content exchange

reward, improved QoS for some time period or various types of discounts in provided network services [30]. However, accepting homogeneity in users' interest may hinder the D2D cooperation opportunities, unless the usage of mobile devices' resources is compensated using assets tailored to users' needs. Therefore, the social context should be enriched with information about users' preferences that can help the operators devise targeted D2D cooperation proposals.

From the users' viewpoint, the introduction of social awareness in D2D cooperative scenarios raises privacy concerns. The acquisition of social characteristics of users in close proximity is of crucial importance in order to identify opportunities for D2D cooperation. Nonetheless, even though this information can help determine trust levels among users, improving the efficiency of D2D cooperative communication, the users might not desire to share personal data about the applications they use or their contact lists. Therefore, their consent to social networking information storing by mobile operators cannot be taken for granted and might be application dependent. For similar reasons, the extent of social trust among users, e.g., trust among friends-of-friends, needs to be properly specified for the formation of trusted cooperative structures. Special attention should be also paid to the design of users' data privacy policies in conjunction with proper encryption methods.

4.6 Chapter concluding remarks

In this chapter, we have highlighted the main challenges of D2D cooperative communication, under the effect of users' social characteristics and the green context of social aware D2D cooperation. We have proposed a social-aware cooperative D2D MAC protocol that promotes the use of friendly users as relays and reduces the energy consumption of D2D cooperation. We also describe some practical concerns that arise when social awareness is incorporated in D2D cooperative networking.

Our simulation results have shown that substantial gains can be achieved if D2D MAC protocols utilize the social information of the cooperating users. More specifically, with SCD2D-MAC, when the density of users belonging to the considered pair's social circle increases, the D2D cooperation potential is reflected in the performance gains. The use of friendly devices and the prioritization of NC-capable relays results in faster data exchange, comparing to the SoA, and SCD2D-MAC achieves a reduction of up to 45% of the D2D data exchange completion time. Additionally, an increase of up to 35% of the energy efficiency is reached and the average battery drain of the mobile devices is up to 58% lower.

In general, even though a social-unaware methodology detects the channel conditions that favor D2D cooperation, it cannot capture the users' social ties. The social structures may be favorable to D2D performance or hinder it, if ignored, as users tend to act altruistically for friends and selfishly for strangers. Additionally, the adaptation of D2D cooperation to the social context can also promote energy awareness, as the existence of social ties might affect the energy consumption levels during cooperation. Therefore,

tackling the challenges of D2D cooperative communication imposes the consideration of the users' intention for cooperation.

Chapter 5

The matching theoretic flow prioritization algorithm

5.1 Introduction

Modern OTT applications can be accessed via Internet connections over cellular networks, possibly shared and managed by multiple MNOs. The OSPs need to interact with MNOs, requesting resources for serving users of different categories and with different QoS requirements. For this purpose, OSPs need OTT application flow prioritization in resource allocation, while the network resource scheduling should respect network neutrality that forbids OSP prioritization. OSPs also need to request resources periodically, according to their performance goals, i.e., GoS level (blocking probability), causing delay in flows' accommodation due to i) the time required for information exchange between OSPs and MNOs, affected by network congestion, and ii) the time required for flows to receive resources, affected by the number of concurrently active flows.

Acknowledging the lack of OSP-oriented resource management approaches and motivated by the aforementioned challenges, in this chapter, we introduce a novel method that allows the intervention of OSPs in the VS allocation in 5G networks. Relying on matching theory, our method enables the OSPs to express interest for resources in eNBs shared by MNOs, aiming to minimize the GoS, without having to inform the MNOs about the exact performance metrics that determine their policies. More specifically, we model the problem as a matching game with contracts [100], where the use of contracts enables the flow prioritization, guaranteeing fairness at the OSP level, as dictated by the network neutrality rules. We define the contract as a combination of parameters that associate a flow with an eNB, indicating the flow's priority and the resources required for achieving

the desired QoS in an eNB. The contracts express the flows' preferences, incorporating the OSPs' policies, and can be ranked by the eNBs in an OSP neutral manner. Additionally, considering that no standard means of interaction between OSPs and MNOs is provided by the current LTE-A specification, we exploit the capabilities of SDN-based network management and use a centralized controller that aggregates the contracts submitted by each OSP independently.

Furthermore, we study the impact of the CN with respect to the congestion levels. Considering the variety of the network topologies and the dynamic nature of the network routes and acknowledging the importance of the RAN in the end-to-end resource allocation, we abstract the CN setup, introducing in our system model the VS allocation step that reflects the CN congestion levels, i.e., higher congestion leads to higher step values. In practice, each step value is induced by the establishment of different routing paths and the allocation of different portions of bandwidth in the CN links. The proposed matching process is repeated in each VS allocation round, thus, the CN congestion levels determine the frequency of the VS allocation process. As the exchanged control messages circulate through the CN nodes, higher congestion in the CN induces higher delay in the transmission of the messages.

In summary, the contribution of this chapter can be described as follows:

(i) *Design of an efficient matching theoretic flow prioritization (MTFP) algorithm:* We first formulate the VS allocation problem incorporating into the mathematical model of matching theory with contracts both the OSPs' policies and the principles of network neutrality that dictate the equal treatment of the different OSPs. Next, we introduce a novel VS allocation algorithm that allows the OSPs to independently i) declare preferences over network resources per VS allocation round and ii) manage their user prioritization policies, respecting the network neutrality with the aid of matching theory and the SDN framework.

(ii) *Description of network architecture that enables the execution of the proposed method:* We present a realistic 4G (and beyond) network architecture that is compliant with the LTE-A specification and employs the SDN framework that enables the proposed algorithm to perform dynamic slicing.

(iii) *Analysis and extensive assessment of the performance of MTFP algorithm in terms of GoS and delay induced by the CN congestion levels:* We design analytical models for the performance evaluation of the MTFP algorithm in terms of GoS and average delay experienced by the flows due the CN impact, considering different OTT application traffic levels and VS allocation frequency, and validate their accuracy through simulations considering various realistic scenarios. Moreover, we assess the performance of the proposed algorithm in terms of achieved GoS, considering different numbers of OTT application flows, and we investigate the experienced delay through extensive simulations.

The remainder of the chapter is structured as follows. In Section 5.2, the considered network architecture and the system model are presented, whereas in Section 5.3, the MTFP algorithm is described. In Section 5.4, a theoretical model of the performance of MTFP algorithm in terms of blocking probability GoS and expected delay experienced by flows that concurrently access a shared RAN is provided. In Section 5.5, we validate the proposed analytical models, investigate the performance of the MTFP algorithm in terms of blocking probability GoS, delay and energy efficiency considering different simulation scenarios and demonstrate the convergence of the MTFP algorithm. We also study the tradeoff between the induced delay and the control overhead of the MTFP algorithm. Last, Section 5.6 concludes this chapter.

5.2 Network architecture and system model

We next describe a shared SDN-based LTE-A network and the system model considered in our study.

5.2.1 Shared SDN–based LTE-A network

In a shared LTE-A network (Fig. 5.1), different MNOs manage cooperatively the RAN elements, e.g., collocated eNBs that cover a geographical area, a pool of RBs and the corresponding CN elements, e.g., switches and routers. The connected UEs use OTT applications of different OSPs. Each application generates data flows that need to be accommodated using end-to-end network resources, i.e., both in the RAN and the CN, allocated in the form of VSs to the corresponding OSPs [115]. Since different OSPs may concurrently claim VSs in the shared network, the VSs should be created in a way that the policies for the flows determined by each OSP are respected, but no prioritization among different OSPs exists according to the network neutrality principle. The implementation of VSs is network specific and can be performed using either of the existing SDN-based solutions for network slicing (e.g., SoftRAN [68], etc.).

In the considered LTE-A network, the network exposure is implemented with the aid of SDN framework, which decouples the control plane from the data plane. The control functions related to RAN and CN entities are managed by logically centralized entities (SDN controllers), whereas the data plane consists of data forwarding elements, e.g., switches and routers, which route the users' flows according to the SDN controllers' instructions [73]. Specifically, an SDN-based virtualization controller (VC) manages three types of software applications that implement functionalities related to CN and RAN control plane: i) the RAN controller (RAN-C) that orchestrates the eNBs, allocating the RBs to flows at each eNB, ii) the core network controller (CN-C) that manages a set of routers, and iii) the OTT services controller (OTTS-C) that is used by the OSPs for OTT service surveillance. For the interaction of MNOs and OSPs with the VC, suitable network APIs are provided. The MNOs access all controllers in the VC through the MNO

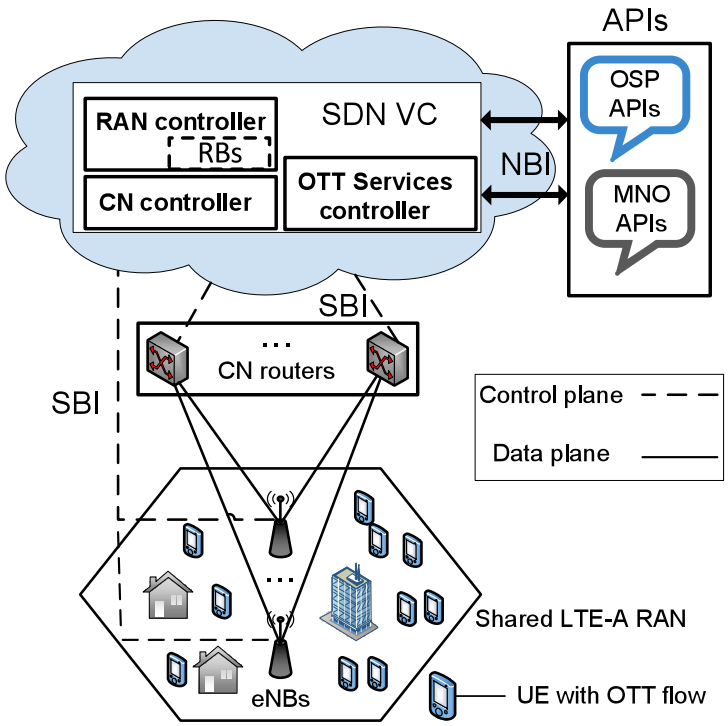


Figure 5.1: Shared SDN–based LTE-A network

API. The OTTS-C communicates with the OSP API and allows the OSPs to assess the flows' performance and request the appropriate resources. The VC can communicate with the eNBs and the routers via a southbound interface (SBI), e.g., OpenFlow, and allows the interaction of the different controllers with the MNO and OSP APIs via a northbound interface (NBI).

In the RAN, the spectrum of each eNB is sliced and shared, thus the VSs offered to OSPs include sets of RBs. Each RB can be allocated only to one eNB in a VS allocation round, thus, the RBs are not re-used in the same cell, avoiding any intra-cell interference issues. In case that neighboring cells share the same pool of resources, inter-cell interference issues may arise, as the same RBs may be re-used, affecting the achievable data rates of UEs in the cell border. In this case, the inter-cell interference coordination (ICIC) mechanism [116] of LTE-A standard can be employed in order to determine disjoint sets of RBs that can be used for the UEs that are affected by inter-cell interference. The resource scheduling is performed periodically, thus, the allocation of RBs to flows is not static throughout the flows' duration and VSs are allocated to OSPs in VS allocation rounds with a frequency determined by the MNOs. The VS allocation frequency allows the transmission of the UEs' information from the shared RAN to the CN and the exchange of the required information between the OSPs and the network resource coordinator. Hence, resource allocation in shared RAN differs from resource scheduling schemes applied in the non-shared network case [17], as a centralized coordinator should divide the resources among the eNBs according to the flows' QoS demands. This process may last longer than the regular resource scheduling performed in every TTI. In the CN,

the aggregation of the flows' information is performed via the available CN links. Thus, when VSs are assigned to OSPs, specific bandwidth is reserved in each CN link.

In order to decide about the VSs needed for the accommodation of the flows' QoS demands, the OSPs should be aware of the status of the UEs related to the flows, e.g., the experienced LTE-A channel conditions. This information is transmitted by the eNBs to the VC. Each UE can connect to an eNB and report its CQI, which determines the MCS used for the downlink transmissions related to the UEs' flows. Thus, the RAN-C can provide the information about flows to OTTS-C, making it available to OSPs' APIs. Using this information, the OSPs' can estimate the QoS levels using the metrics they prefer and adjust their policies, i.e., requirements regarding the allocated VSs.

5.2.2 System model

We consider the cell of a shared RAN jointly operated by N MNOs that have deployed collocated eNBs (Fig. 5.2). Each MNO owns an eNB $n \in \mathcal{N}$ and spectrum, both shared with the other MNOs. A resource pool of W RBs is available, whereas U UEs are connected to the network as subscribers of either of the MNOs. A set of $\mathcal{M}$ OSPs co-exist in the network and each UE may generate flows related to different OTT applications. Thus, each flow corresponds to a specific UE and OSP. Assuming a set of $\mathcal{J}$ OTT application flows of different OSPs and m a specific OSP, we denote $\mathcal{J}^{(m)}$ the set of flows related to the OTT application of OSP m.

The OSPs have policies for the OTT service differentiation that determine the flows' importance in the VS allocation process. It should be noted that the OTT service differentiation does not affect network neutrality, as it refers to the internal policies of the OSPs. Thus, the flows have different characteristics and different user priorities exist. Each flow's priority p_j is set by the OSP. Flows of different OTT applications may have different priorities, even when the flows are related to the same UE. The downlink traffic flows related to the OTT applications are generated by U UEs following a Poisson distribution with rate λ (flows/hour/UE) [1]. Given a set of $\mathcal{K}$ priority classes, we denote by $\lambda_{k,m}$ the flow generation rate per priority class k for each OSP $m \in \mathcal{M}$. The duration of each flow is exponentially distributed with mean equal to $1/\mu$. Each OSP needs to acquire a number of RBs in order to serve the flows associated with UEs in either of the available eNBs. The VC virtualizes the eNBs and the spectrum in a way that v_m RBs are allocated to the VS that corresponds to OSP $m \in \mathcal{M}$. Each flow $j \in \mathcal{J}^{(m)} \subset \mathcal{J}$ needs a number of $v_n^{(m,j)} \leq v_m$ RBs that provides it with a downlink data rate $r_{\text{srv}}^{(m,j)}$ [2].

As each flow is associated to a specific UE, the downlink channel status is reported to the VC in order to enable the OSPs to decide upon the resources that are requested

[1] In order to study theoretically the performance of the proposed algorithm, we use the Poisson traffic model that is commonly used to model voice sessions and is also suitable for the scenario of video streaming sessions ([117], [118], etc.). The use of a different traffic generation model would not affect the problem formulation, the functionality of the proposed algorithm and the main conclusions of our study.

[2] Uplink traffic flows could be also considered.

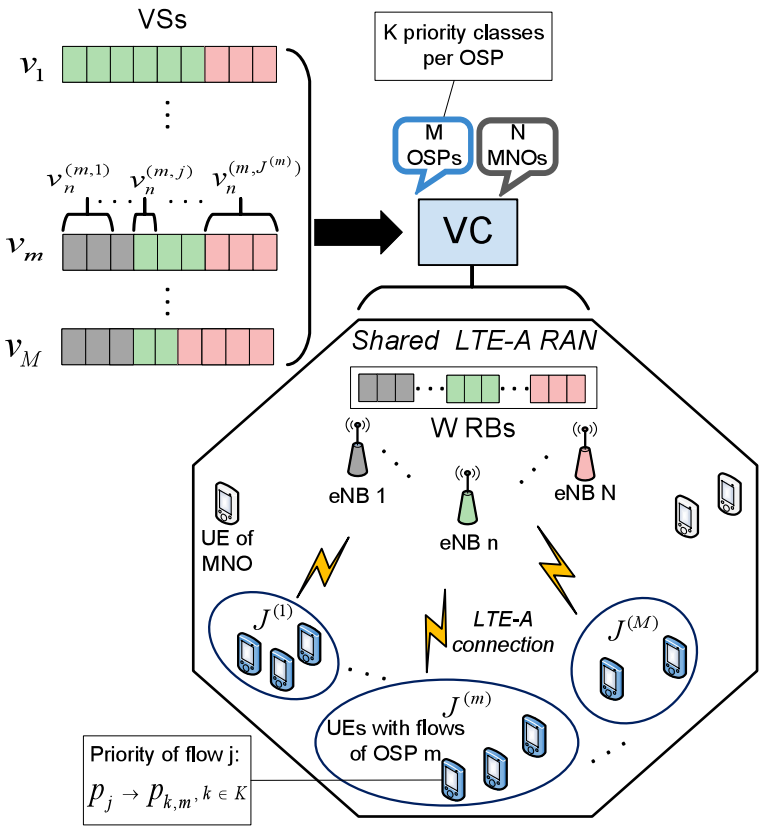


Figure 5.2: VS allocation in the considered network

per VS allocation round. In the considered shared network, a UE that generates a flow can report CQIs for each eNB n in every TTI [17]. Given an $\text{MCS}_n^{(m,j)}$ and a number of allocated RBs $v_n^{(m,j)}$ to the UE related to flow j, the achievable downlink data rate is estimated as:

$$r_n^{(m,j)} = \frac{L\left(\text{MCS}_n^{(m,j)}, v_n^{(m,j)}\right)}{\text{TTI}}, \tag{5.1}$$

where $L(\text{MCS}_n^{(m,j)}, v_n^{(m,j)})$ is the transport block size [106]. The value $\text{MCS}_n^{(m,j)}$ may be different in each round for a specific UE. Moreover, each UE experiences different SNR levels, thus different MCS values are reported. We assume downlink channels with Rayleigh fading, such that the SNR can be represented by a random variable with average value γ and probability density function given by:

$$f(x) = \frac{1}{\gamma} e^{-\frac{x}{\gamma}} u(x), \tag{5.2}$$

where $u(x)$ is the unit step function. The probability ρ_i that the ith MCS is selected out of the set $\mathcal{I}$ of possible MCSs can be derived as:

$$\rho_i = \int_{\gamma_{thr}^{(i)}}^{\gamma_{thr}^{(i+1)}} f(x)dx = e^{\frac{\gamma_{thr}^{(i+1)}}{\gamma}} - e^{\frac{\gamma_{thr}^{(i)}}{\gamma}}, \tag{5.3}$$

where γ is the average SNR and $[\gamma_{thr}^{(i)}, \gamma_{thr}^{(i+1)}]$ denotes the SNR range that corresponds to

MCS i. The SNR of each UE varies randomly in each VS allocation round [3].

As explained in Section 5.2.1, the VS allocation and assignment of RBs to flows is performed periodically in successive VS allocation rounds. The OSPs request RBs with step t, which is a random variable exponentially distributed with mean value $\mathbb{E}[t] = 1/\nu$, lower bounded by the time required for the UEs' CQIs to be sent to the VC. While a UE that generates a flow j maintains the connection to the corresponding OTT application active, the flow experiences several rounds. However, in each round, RBs may or may not be allocated to a flow j, as it should hold that $\sum_{m \in \mathcal{M}} v_m \leq W$. Thus, a flow j experiences a delay d_j, related to the time spent in fruitless rounds and the average experienced delay of all flows is defined as $\mathbb{E}[D]$.

In each VS allocation round, control messages are exchanged between RAN and VC for the coordination of VS allocation. The exchange of control messages occupies bandwidth in the CN links that comprise the paths from RAN to VC, increasing the control overhead β, i.e., the ratio of the size s_{ctrl} of the control messages sent through the CN links over the total size of useful data s_{data} sent per round (OTT application data packets sent to UEs) and the size s_{ctrl}:

$$\beta(\%) = \frac{s_{ctrl}}{(s_{data} + s_{ctrl})} 100. \tag{5.4}$$

Lower ratio β implies lower overhead per round. The total size of data sent per round is $s_{data} = r_e \mathbb{E}[t]$, where $\mathbb{E}[t]$ is an average VS allocation step value and r_e is the effective throughput in the RAN-VC path. The value r_e is affected by the network topology, e.g., when multihop paths from RAN to VC exist, it is bounded by the minimum of the data rates at each hop [119].

The network energy efficiency is affected by the total data rate demand in each eNB, i.e., the number of served flows and their data rate requirements, and the channel conditions of the UEs, i.e., the total number of RBs used by the corresponding eNBs [4]. Using Eq. (5.1), we define the energy efficiency η_n per eNB n in a VS allocation round as:

$$\eta_n = \frac{\sum_{m \in \mathcal{M}} \sum_{j \in \mathcal{J}^{(m)} \cap \mathcal{J}_{(n)}} r_n^{(m,j)}}{P_n}, \tag{5.5}$$

where the power consumption P_n of eNB n is equal to [120]:

$$P_n = P_C^{(n)} + \delta P_{RB}^{(n)}, \tag{5.6}$$

considering, for each eNB n, the constant power consumption $P_C^{(n)}$ related to signal processing, cooling and battery backup, the power consumption δ that scales with the average radiated power due to amplifier and feeder losses and the power consumption $P_{RB}^{(n)}$ for the transmission of one RB. Given W available RBs, the transmission power $P_{Tx}^{(n)}$ and a_n the number of antennas of eNB n, the value $P_{RB}^{(n)}$ is calculated as:

$$P_{RB}^{(n)} = \frac{P_{Tx}^{(n)}}{a_n W}. \tag{5.7}$$

[3] The value ρ_i is only required for the GoS analysis presented in Section 5.4.1 and the delay analysis presented in Section 5.4.2.

[4] We assume that no capacity or power constraints are applied for the eNBs.

Using Eq. (5.5), we derive the overall network efficiency as:

$$\mathbb{E}\left[\eta\right] = \frac{\sum_{n\in\mathcal{N}} \eta_n}{|\mathcal{N}|}, \tag{5.8}$$

assuming a number of $\mathcal{N}$ eNBs in the shared network.

5.3 Matching theoretic flow prioritization

In this section, we describe the VS allocation problem for OSPs and propose a flow prioritization scheme that relies on matching theory.

5.3.1 VS allocation and involved parties' preferences

In a shared RAN, different resource allocation policies can be employed, based on well-known scheduling techniques, e.g., round robin or maximum throughput scheduling, which achieve different performance goals of MNOs [17]. When the OSPs' preferences have to be considered, the flows' priorities should be taken into account in each VS allocation round in a way that flows of higher priority receive resources first.

The process of VS allocation to OSPs involves the assignment of RBs to flows according to two types of parameters: i) network-related parameters, i.e., current CQI and MCS values of the UE related to a flow, monitored by the VC, and ii) application-related parameters set by the OSPs, i.e., required QoS levels (minimum acceptable data rate), and flow priority defined by the corresponding OSP's policy. At each VS allocation round, each OSP m seeks to obtain RBs in the eNBs that offer the requested downlink data rates $\sum_{j\in\mathcal{J}^{(m)}} r_{\text{srv}}^{(m,j)}$, with respect to the flows' priorities, and minimize the blocking probability GoS_m, i.e., the ratio of the number of flows that are not served with the required data rates over the total number of flows $\mathcal{J}^{(m)}$:

$$GoS_m = 1 - \frac{1}{|\mathcal{J}^{(m)}|} \sum_{j\in\mathcal{J}^{(m)}} \sum_{n\in\mathcal{N}} [r_n^{(m,j)}(v_n^{(m,j)}) \geq r_{\text{srv}}^{(m,j)}] \in [0,1], \tag{5.9}$$

Let us recall that the allocation of RBs may not be possible for all flows at each VS allocation round. Each OSP prefers that flows with higher priority, i.e., lower p_j value, receive the required RBs first in each VS allocation round, ensuring that they experience lower delay than flows of lower priority. Among flows with the same priority, those that have lower demands of RBs, e.g., experience better channel conditions or have lower data rate demands, should be served first.

The MNOs aim to minimize the expected number of flows of all OSPs that do not achieve the required data rates, i.e., the $\mathbb{E}\left[GoS\right]$, respecting the OSPs' priorities without violating the network neutrality. The value $\mathbb{E}\left[GoS\right]$ is equal to:

$$\mathbb{E}\left[GoS\right] = \frac{\sum_{m\in\mathcal{M}} GoS_m}{|\mathcal{M}|}. \tag{5.10}$$

To guarantee network neutrality, two conditions should hold: (a) there should exist at least one flow $j \in \mathcal{J}^{(m)}$ and at least one flow $j' \in \mathcal{J}^{(m')}$, such that $p_j = p_{j'}$ and $d_j > d'_j$, and (b) there should exist at least one flow $j'' \in \mathcal{J}^{(m)}$ such that $p'_j = p_{j''}$ and $d'_j < d''_j$. The conditions (a) and (b) state that no OSP should gain priority over the other OSPs, achieving delay for its flows that is lower than the delay experienced by the flows of the same priority class of the other OSPs. It should be noted that, when the OSPs' policies are considered, flows of lower priorities may be lead to starvation, as the spectrum capacity may not be sufficient. Therefore, the eNBs can update the priorities submitted by the OSPs depending on whether each flow has previously received resources or not, in order to both respect the OSPs' policies and guarantee that all flows receive resources at some point. The higher the priority of a flow, the more likely it is that it receives resources at a VS allocation round and the lower is the experienced delay.

5.3.2 Formulation of matching process using contracts

We thereupon provide the necessary matching-theoretic definitions that describe the concepts employed by the proposed OTT flow prioritization approach (Section 5.3.3).

The VS allocation process resembles the hospital-doctor matching problem [100], where doctors seek to be matched with hospitals, achieving the highest possible wage or better working conditions. In the considered problem, the flows offer contracts, whereas eNBs act as the hospitals that rank the offered contracts. In our work, we define the contract as a combination of parameters that associate a flow with an eNB, i.e., it contains the flow's priority and the RBs required for achieving the desired QoS in a specific eNB. A flow must be associated with exactly one eNB and an eNB can serve multiple flows (many-to-one matching). For each flow there exist several possible contracts that may be preferable. It is also possible that a flow will not obtain any contract, thus it will not be allocated resources in any eNB, accepting a null contract.

5.3.2.1 Definition of contracts and preferences of players

A contract c related to flow j and eNB n is represented by a vector (j, n, q), where q is the cost of contract $q = p_j.v_n^{(m,j)}$ that is defined as a real number with the integer part equal to the flow's priority p_j and a decimal part equal to the RBs $v_n^{(m,j)}$ required by the UE related to flow j in order to achieve $r_{\text{srv}}^{(m,j)}$, when the UE is connected to eNB n, as given by Eq. (5.1).

The flows create a preference list of $(|\mathcal{K}||\mathcal{N}| + 1)$ contracts with cost values q that denote the most preferred priority and RBs per eNB, including the null contract. The lower the value p_j the higher the priority of the flow, e.g., a high priority flow has a value $p_j = 1$, which denotes higher priority than a flow with $p_j = 2$ and increases its chances of receiving RBs reducing the experienced delay. The term $v_n^{(m,j)}$ can take any value from one to the maximum number of RBs that can be assigned to a UE [121]. Let us now consider an example with two eNBs and a flow with high priority ($p_j = 1$) that can be

served with the requested data rate occupying 3 RBs in the first eNB and 5 RBs in the second eNB. The contracts with q values $(1.3, 1.5)$ are the most preferred, as they denote the desired priority. In order to avoid staying unmatched in case that an eNB prefers other flows of high priority, the flow also includes two contracts in the preference list that denote the next lowest priority, i.e., $(2.3, 2.5)$ and the contracts in the list of the flow are ordered as $(1.3, 1.5, 2.3, 2.5, \varnothing)$.

Therefore, a preference relation of a flow $j \in \mathcal{J}$ over the available eNBs $n \in \mathcal{N}$ is a relation over the set of the available contracts, including the null contract, which implies that no association exists between an eNB and a flow. For a flow j, we define a preference relation $\succeq_j$ over the set of contracts C such that for any two contracts $c', c'' \in C$ with costs q' and q'', respectively, the flow prefers the contract with the lower cost, thus, the preference relation can be defined as:

$$c' \succeq_j c'' \Leftrightarrow q' \geq q''. \tag{5.11}$$

The rationale of each eNB's preferences is similar, as it also prefers the contracts with the minimum possible cost and it additionally takes into account whether a specific flow has been served in the previous VS allocation round, in order to guarantee all flows receive resources at some point. In our study, we assume that the eNBs are operated by MNOs that have the same performance goal, i.e., minimize the GoS. However, the eNBs may also have different preferences, expressing different objectives of the MNOs for the network performance.

Let us now denote by τ a round, $\tau + 1$ the next round and the set of served flows in a specific eNB n in round τ as $S_n^{srv}(\tau)$. Assuming that two contracts c' and c'' appear in round $\tau + 1$ and are submitted by flows j and j', respectively, which have the same priority, i.e., $p_{j'} = p_{j''}$. If flow j'' has been previously served by the same eNB, i.e., it belongs to the set $S_n^{srv}(\tau)$ and flow j' has not been served by eNB n, then contract c' is preferred. Therefore, we can define the preference relation $\succeq_n$ of an eNB n over the set of contracts C in a round $\tau + 1$ as follows:

$$c' \succeq_n c'' \Leftrightarrow j' \notin S_n^{srv}(\tau)'' \text{ and } j'' \in S_n^{srv}(\tau) \text{ and } p_{j'} = p_{j''}. \tag{5.12}$$

5.3.2.2 Properties of stable matching

We next describe the properties used in order to characterize the flow-eNB association as stable. The contracts that are accepted confirm the agreement between flows and eNBs and form the chosen set, whereas the rest of the contracts form the rejected set. Letting $\mathcal{N}$ be the set of eNBs, $\mathcal{J}$ the set of OTT application flows and Q the set of all possible costs, the set of all possible contracts C is defined as $C = \mathcal{J} \times \mathcal{N} \times Q$ [122].

Definition 2. *Given the set of all possible contracts C and $C' \subset C$ a subset of C, the chosen set $S_j(C')$ of a flow j either contains only one element (the flow's preferred contract out of C') or is empty, if there is no acceptable contract c in C' for flow j. Similarly, the chosen set $S_n(C')$ of an eNB n either contains the eNB's preferred contracts out of C' or is empty, if there is no acceptable contract c in C' for eNB n.*

The remaining contracts that are not accepted from anyone form the set of rejected contracts.

Definition 3. *Given the set of all possible contracts C, a subset C' of C, and $S_J(C') = \cup_{j\in\mathcal{J}} S_j(C')$ and $S_N(C') = \cup_{n\in\mathcal{N}} S_n(C')$ the chosen sets of all flows and eNBs, respectively, the sets of contracts that are rejected by all flows and all eNBs are defined as $R_F(C') = C'\backslash S_J(C')$ and $R_N(C') = C'\backslash S_N(C')$. The rejected sets of a flow j and an eNB n are defined as $R_j(C')$ and $R_n(C')$, respectively.*

A stable association between eNBs and flows is achieved, if there exists no allocation strictly preferred by any eNB and weakly preferred by all flows related to a specific eNB, and there exists no flow that would prefer to reject the contract it has received. An allocation is weakly preferred by a flow if the flow desires it at least as much as any other allocation.

Definition 4. *A set of contracts $C' \subset C$ results in a stable VS allocation if and only if*

(i) $S_N(C') = S_J(C') = C'$ *(*individual rationality*)*

*(ii) there exists no eNB $n \in \mathcal{N}$ and set of contracts $C'' \neq S_n(C')$ such that $C'' = S_n(C' \cup C'') \subset S_J(C' \cup C'')$ (*nonexistence of blocking contracts*).*

The first condition dictates that if only the contracts in C' are available, then they are all chosen. When the condition does not hold, it means that there exist a flow or eNB that prefers to reject a contract. According to the second condition, there exist no set of contracts C'' that could be added and would be selected by both eNB n and the flows related to n. Thus, the matching is not blocked by any flow or eNB.

It has been proven that the property of substitutability for the eNBs' preferences is a sufficient condition for achieving a stable allocation [100].

Definition 5. *The contracts in C are considered to be substitutes for any eNB $n \in \mathcal{N}$, if for all subsets $C' \subset C'' \subset C$, it holds that $R_n(C') \subset R_n(C'')$, where R_n is the set of contracts rejected by n, i.e., the rejection sets $R_n(C')$ and $R_n(C'')$ are isotone. (*substitutability*).*

According to the property of substitutability of eNBs' preferences over contracts, every contract rejected from C' is also rejected from C'', and if a contract is chosen by an eNB from some available contracts, then that contract will still be selected from any smaller set that includes it. Thus, the contracts of an eNB n are substitutes, if for any contracts $c', c'' \in C$ and any sets $C' \subset C$, it holds that $c'' \in S_n(C' \cup \{c', c''\}) \Rightarrow c'' \in S_n(C' \cup \{c''\})$.

5.3.3 Proposed matching theoretic approach

We next present the MTFP algorithm that matches the flows that access a shared LTE-A network, considering their priorities, with the eNBs. The proposed algorithm is based

on the matching process presented in [100] and describes the way the players interact with each other in practice, i.e., how the submission of contracts is performed. The VS allocation process is repeated periodically, thus, the MTFP algorithm is applied in each VS allocation round. The MTFP algorithm is described in Section 5.3.3.1. The MTFP control overhead and complexity are discussed in Section 5.3.3.2 and the exchange of control messages is detailed in Section 5.3.3.3.

5.3.3.1 Description of the MTFP algorithm

Algorithm 1 consists of two phases (i.e., initialization and negotiation) that are performed in each VS allocation round. The initialization phase refers to the collection of flows' information and the OSPs' requirements by the VC. In the negotiation phase, the matching process is performed by the VC that is an entity trusted by the OSPs and is fundamental for the implementation of MTFP, as the OTTS-C is the entity that interacts with the various OSP APIs via the exchange of control messages. Given that no standard means of interaction between OSPs and MNOs is provided by the current LTE-A specification, with the VC, we exploit the capability of centralized network management offered by the framework of SDN.

In the initialization phase, all UEs report their CQIs and the eNBs transmit this information to the VC (in RAN-C). The OSPs update the information about the priorities of their flows and the required QoS. In the negotiation phase, at each matching iteration, the flows rank their contracts, according to the priorities set by their OSPs, and submit their most preferred contracts to the corresponding eNBs via the OTTS-C. The eNBs update in RAN-C the flows' priorities and sort the available contracts. Two sets of contracts are next created, i.e., the chosen set S_N that contains the most preferred contracts from the flows' perspective based on the OSPs' preferences and the rejected set R_N, which is the complement of the chosen set. The negotiation phase is repeated while the rejected flows submit requests for assignment to their next preferred set of contracts, until no more contracts are added to the rejected set R_N. Once contracts are finalized, the requested RBs are allocated to the eNBs and the VSs are created. The MTFP algorithm is applicable independently of the slice isolation technique employed by the VC, as it does not intervene to the implementation of the VSs. With the dynamic slicing that it performs, isolation is maintained, as each RB is assigned at most to one eNB per VS allocation round. The CN resources are allocated to the flows according to the RB allocation.

Proposition 2. The MTFP algorithm converges to a stable eNB-flow matching through contracts after a finite number of iterations.

Proof. The MTFP algorithm is based on the matching process presented in [100] that addresses the hospital-doctor association problem. Therefore, the iterations stop and the algorithm converges when no more flows are added to R_N, thus, every flow is associated with an eNB and the property of substitutability (Definition 1) characterizes the eNBs' preferences. ∎

Algorithm 1 Matching theoretic flow prioritization (MTFP) algorithm

Input: CQIs of UEs, rate constraints and priorities of flows

Output: Stable allocation per VS allocation round

Initialization phase:

1: The UEs with active flows submit their CQIs to eNBs.

2: The eNBs submit the flows' information to VC.

3: Each OSP m checks each flow's j status and assigns the priorities p_j and requested data rate $r_{\text{srv}}^{(m,j)}$.

Negotiation phase: // Start matching iterations

4: **Repeat:**

5: The flows estimate the RBs required at each eNB n and sort the available contracts $c \in C$ according to cost $q \in Q$.

6: Each flow (in OTTS-C) $j \in \mathcal{J}$ creates the chosen set $S_j(C')$ and the rejected set $R_j(C') = C' \backslash S_j(C'), C' \subset C$.

7: Each eNB (in RAN-C) $n \in \mathcal{N}$ updates the priorities of flows that have been served in previous VS allocation round (p_j =initial $p_j + 1$).

8: Each flow with $R_j(C') \neq \emptyset$ submits the next preferred contract from $S_j(C')$ to the VC.

9: The eNBs check if the flows that submit contracts have been previously served:

10: $\forall$ flow $j \in \mathcal{J}$:

11: **if** flow j rejected in the previous VS allocation round **then**

12: Set p_j = initial p_j.

13: **end if**

14: Each eNB n accepts most preferred contracts out of those offered in the current iteration and rejects the others, creating the chosen set $S_n(C')$ and the rejected set $R_n(C') = C' \backslash S_n(C'), C' \subset C$.

Until convergence to a stable allocation.

15: The VC assigns RBs to flows considering the number of available RBs W and transmits the required information to the eNBs.

Table 5.1: Contracts submitted by OTT application flows

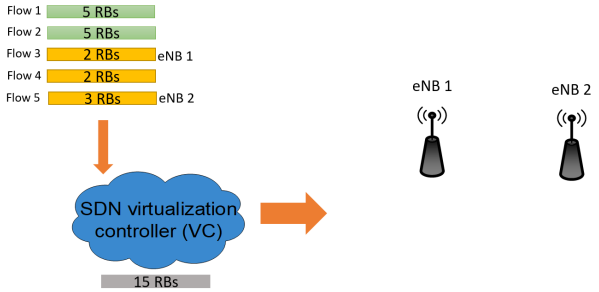
Flow ID	1	2	3	4	5
contract=(eNB, priority, RBs)	(1,1,5) (2,1,6) (1,2,5) (2,2,6) null	(1,1,5) (2,1,6) (1,2,5) (2,2,6) null	(1,2,2) (2,2,3) (1,3,2) (2,3,3) null	(1,2,2) (2,2,3) (1,3,2) (2,3,3) null	(2,2,3) (1,2,6) (2,3,3) (1,3,6) null

Let us know provide a simple operation example of MTFP algorithm that demonstrates the matching process in a VS allocation round. In Fig. 5.3, the iterations performed until the MTFP algorithm converges to a stable matching are depicted, considering the lists of the preferred contracts of the flows shown in Table 5.1. We assume that two eNBs and a total number of 15 RBs are available in the network. In a specific VS allocation round, two flows (flow 1 and 2) of high priority appear for the first time, each requiring 5 RBs from eNB 1, and three flows (flows 3-5) of low priority also request RBs. Each of the flows 3 and 4 require 2 RBs in eNB 1 and flow 5 requires 3 RBs in eNB 2, whereas flows 3 and 5 have been previously served by eNBs 1 and 2, respectively. The flows submit their most preferred contracts in iteration 1 and the eNBs respond to the flows' requests. Flows 1, 2 and 4 are accepted by eNB 1, whereas flows 3 and 5 are rejected by the eNBs they prefer (eNB 1 and 2, respectively) because they have been served by them in the previous round. In iteration 2, the rejected flows submit the next contracts in their list and flow 3 is accepted by eNB. Flow 5 is rejected, as no more RBs are available, and continues to submit contracts until no more options in its list exist. The matching is completed after 5 iterations.

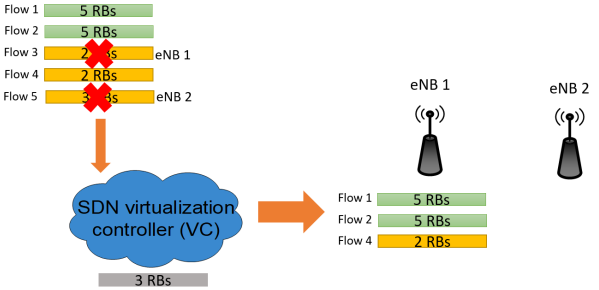
5.3.3.2 Overhead and complexity of MTFP algorithm

We next discuss the overhead induced by the exchange of control messages and the computational cost of the MTFP algorithm.

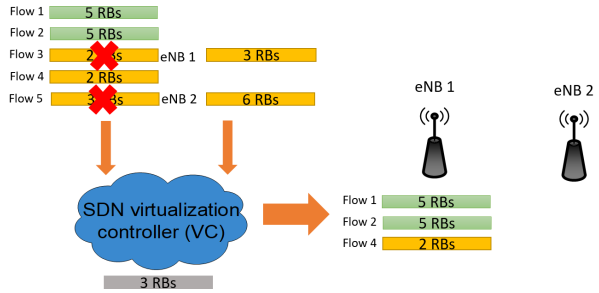
Regarding the induced control overhead, control messages are exchanged in both phases of MTFP (Section 5.3.3). In the initialization phase, a number of U UEs that concurrently need resources for their flows report their CQIs to $|\mathcal{N}|$ eNBs, transmitting $O(U|\mathcal{N}|)$ messages. The eNBs transmit the CQIs to the VC, thus $O(|\mathcal{N}|)$ messages are sent. In the negotiation phase, the matching process requires the exchange of messages among the OTTS-C and the RAN-C. Flows and eNBs exchange contract proposals through the VC, until every flow is associated with an eNB. Considering that $|\mathcal{N}|$ eNBs and $|\mathcal{K}|$ priority classes exist, a number of $(|\mathcal{K}||\mathcal{N}|+1)$ possible contracts are provided. Instead of sending one message for each flow's proposal, one message can be sent on behalf of each OSP, containing the proposals of the related flows. Assuming the worst case that would require all flows to submit all the available proposals before being matched, at most $O(|\mathcal{M}|(|\mathcal{K}||\mathcal{N}|+1))$ messages are sent from OTTS-C to the RAN-C and vice



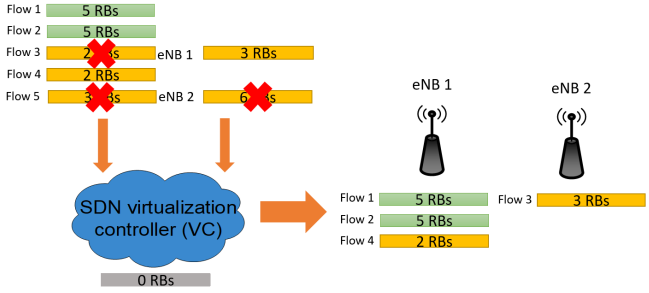
(a) Iteration 1: submission of contracts



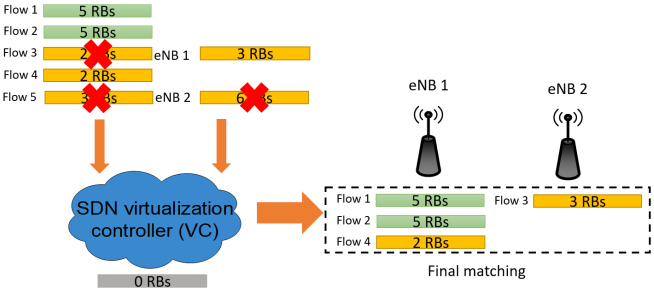
(b) Iteration 1: responses to contracts



(c) Iteration 2: submission of contracts



(d) Iteration 2: responses to contracts



(e) Final matching in iteration 5

Figure 5.3: Operation example of MTFP algorithm

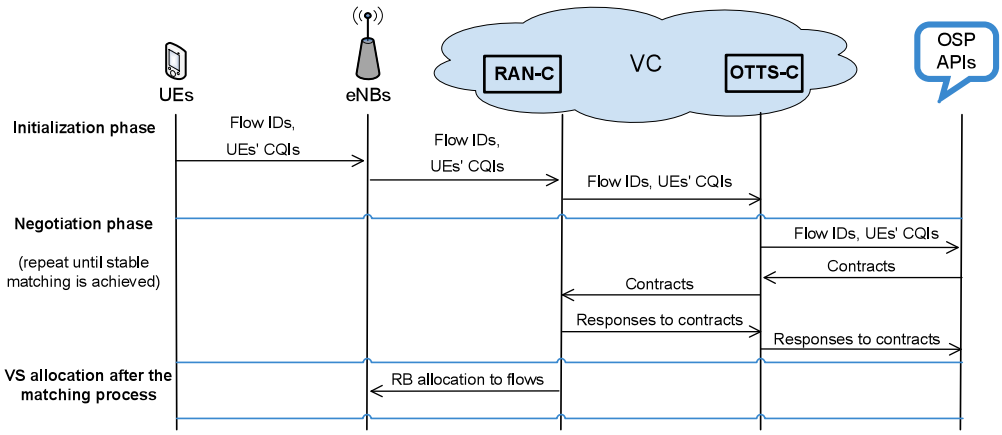


Figure 5.4: Messages exchanged for the application of the MTFP algorithm

versa, considering that $|\mathcal{M}|$ OSPs exist. Finally, after the matching process is completed, the VC informs the eNBs about the RBs that should be allocated to the flows, sending $O(|\mathcal{N}|)$ messages.

The computational complexity is related to the sorting operation in the negotiation phase. Assuming $|\mathcal{J}|$ flows, at each iteration, each flow sorts a list of $(|\mathcal{K}||\mathcal{N}|+1)$ elements, inducing a total complexity of $O((|\mathcal{K}||\mathcal{N}|+1)\log(|\mathcal{K}||\mathcal{N}|+1))$. Similarly, given $|\mathcal{N}|$ eNBs, each sorting a list of $|\mathcal{K}||\mathcal{J}|$ elements, the complexity of the sorting operation is equal to $O(|\mathcal{K}||\mathcal{J}|\log(|\mathcal{K}||\mathcal{J}|))$. As $|\mathcal{M}|$, $|\mathcal{N}|$ and $|\mathcal{K}|$ are much smaller than $|\mathcal{J}|$, the resulting complexity is $O(|\mathcal{J}|\log|\mathcal{J}|)$.

Overall, the practicality of MTFP algorithm is mostly affected by the exchange of control messages, which increase proportionally to the number U of active UEs.

5.3.3.3 Control messages in MTFP algorithm

For the application of MTFP (Section 5.3.3.1), control messages are exchanged in a VS allocation round (Fig. 5.4). In the initialization phase of MTFP, each connected UE reports the flow ID and CQI to each eNB by sending a control message (step 1). In step 2, each eNB aggregates the IDs and CQIs received by the UEs and sends a message containing the IDs and the CQIs to the RAN-C in the VC. At step 3, the RAN-C sends one message with this information to the OTTS-C, which next communicates the information to the OSPs, sending to each OSP a message containing the information of the flows that are related to the specific OSP. During the matching process that is performed in steps 4 to 14, each flow submits the most preferred contract to the OTTS-C (step 9). As several flows may belong to the same OSP, their contracts are contained in a single message originating from the OSP API, which is sent to the OTTS-C and forwarded to the RAN-C. The RAN-C next communicates the decision about the flows' contracts, sending a message to the OTTS-C, which next notifies the OSPs about the decision sending to each OSP API one message (step 14). The transmission of messages containing contracts and responses to contracts continues until a stable matching is achieved. After the negotiation

phase ends, the RAN-C notifies the eNBs about the RB allocation sending a message to each eNB (step 15).

5.4 Performance analysis of MTFP algorithm

In this section, we provide a theoretical model of the performance of MTFP algorithm in terms of blocking probability GoS and expected delay experienced by flows that concurrently access a shared RAN.

As discussed in Section 5.3.3, MTFP creates VSs for the OSPs' flows by repeating periodically a matching process. In each VS allocation round, the number of flows served with the requested GoS is limited by the number of available RBs. Moreover, the flows that have not received resources in one round may be served in a subsequent round. Thus, a flow experiences time delay until it obtains RBs. The average delay a flow is expected to experience during several rounds is affected by the network status, i.e, the flow generation rate, the mean flow duration, the number of priority classes of flows that coexist, the flows' QoS demands in terms of data rate and the number of available RBs. Considering these parameters, we analytically derive the blocking probability GoS in each round and the expected delay when MTFP is applied.

5.4.1 GoS analysis

Let us now consider the shared network of Fig. 5.2 that serves $|\mathcal{J}^{(m)}|, m \in \mathcal{M}$ flows at a specific VS allocation round. The OSPs related to the flows share W RBs and each flow j related to OSP m requires a specific number of RBs in order to be served with data rate $r_{\text{srv}}^{(m,j)}$. Considering downlink channels with Rayleigh fading and different rates $r_{\text{srv}}^{(m,j)}$ required by the flows, the expected total number of RBs $\mathbb{E}[b_T]$ needed by all flows can be estimated using Eq. (5.3) as:

$$\mathbb{E}[b_T] = \sum_{m \in \mathcal{M}} \sum_{j \in \mathcal{J}^{(m)}} \sum_{i \in \mathcal{I}} \rho_i \phi(i, r_{\text{srv}}^{(m,j)}), \tag{5.13}$$

where ϕ is a function that searches the table reported in [106] and returns the minimum transport block size that can be used in order that the flow achieves the requested data rate with MCS i. Given Eq. (5.13), the expected blocking probability GoS can be calculated as follows:

$$\mathbb{E}[GoS] = \begin{cases} 0, \text{if } W > \mathbb{E}[b_T] \\ 1 - \left\lceil \dfrac{W}{\mathbb{E}[b_T]} \right\rceil, \text{if } W \leq \mathbb{E}[b_T], \end{cases} \tag{5.14}$$

when a number of W RBs is available in the shared RAN.

5.4.2 Delay analysis

In the network depicted in Fig. 5.2, OTT application flows are generated by U UEs, with rate λ from each UE, thus the total rate is $U\lambda$. Each flow needs an average number of $\mathbb{E}[b]$ RBs per VS allocation round. Assuming $|\mathcal{K}|$ priority classes per OSP and considering Eq. (5.3), $\mathbb{E}[b]$ is equal to:

$$\mathbb{E}[b] = \sum_{i \in \mathcal{I}} \rho_i \phi(i, \mathbb{E}[r_{\text{srv}}]), \tag{5.15}$$

where $\mathbb{E}[r_{\text{srv}}]$ is the average required data rate weighted by the coefficients $\lambda_{k,m}$, i.e., the flow generation rate per priority class $k \in \mathcal{K}$ for each OSP $m \in \mathcal{M}$. The value $\mathbb{E}[r_{\text{srv}}]$ is estimated as the following weighted average:

$$\mathbb{E}[r_{\text{srv}}] = \frac{\sum_{k \in \mathcal{K}} \sum_{m \in \mathcal{M}} \lambda_{k,m} r_{\text{srv}}^{(k,m)}}{\sum_{k \in \mathcal{K}} \sum_{m \in \mathcal{M}} \lambda_{k,m}}, \tag{5.16}$$

where $r_{\text{srv}}^{(k,m)}$ is the data rate required by priority class k flows per OSP m.

For the accommodation of a flow, a set of $\mathbb{E}[b]$ RBs, defined as cluster, is required. Considering that W corresponds to the number of the total available RBs, the number of clusters that exist in the system can be defined as $X = \lceil W/\mathbb{E}[b] \rceil$. If the clusters cannot serve all the active flows, the flows that have not received RBs join a queue (orbit queue), with maximum capacity $Y = W$, and wait until they are served. Each flow aims to occupy a cluster for a service exponentially distributed with mean equal to $1/\mu$. Furthermore, every flow in the orbit queue can request resources in the round. Hence, we can view the network as a finite source retrial queuing system where the retrial rate is exponentially distributed with mean value equal to $1/\nu$.

We model the considered system using a continuous time Markov chain (CTMC) with state space $\mathbb{A} = \{(x, y) | 0 \leq x \leq X, 0 \leq y \leq Y\}$, where x is the number of occupied clusters and y the number of flows in the orbit queue, which define a system state (x, y). The flows experience an average delay $\mathbb{E}[D]$. We denote as $\mathbb{E}[X]$ the average number of occupied clusters and as $\mathbb{E}[Y]$ the orbit queue length. Considering Little's Law [123], we derive the following equation:

$$\mathbb{E}[D] = \frac{\mathbb{E}[Y]}{\mathbb{E}[\lambda]}, \tag{5.17}$$

where $\mathbb{E}[\lambda]$ is the expected flow arrival rate at the network, including new flows that are generated and flows that reside in the orbit queue. Given that the utilization ratio of the clusters is equal to $\mathbb{E}[X]/X$ and the average time a flow aims to reside in the cluster is $1/\mu$, we observe that $\mathbb{E}[\lambda] = \mathbb{E}[X]/(1/\mu) = \mathbb{E}[X]\mu$. For the calculation of the expected delay $\mathbb{E}[D]$, the values $\mathbb{E}[X]$ and $\mathbb{E}[Y]$ have to be estimated.

The considered network is a CTMC that can be described by the steady state probabilities $\pi(x, y)$, as shown in Fig. 5.5. Each horizontal line of the diagram refers to transitions between states that refer to the same orbit queue length but different number of occupied clusters. New flows that arrive and are served increase the number of occupied clusters, whereas this number reduces when flows leave the system. The diagonal lines denote the

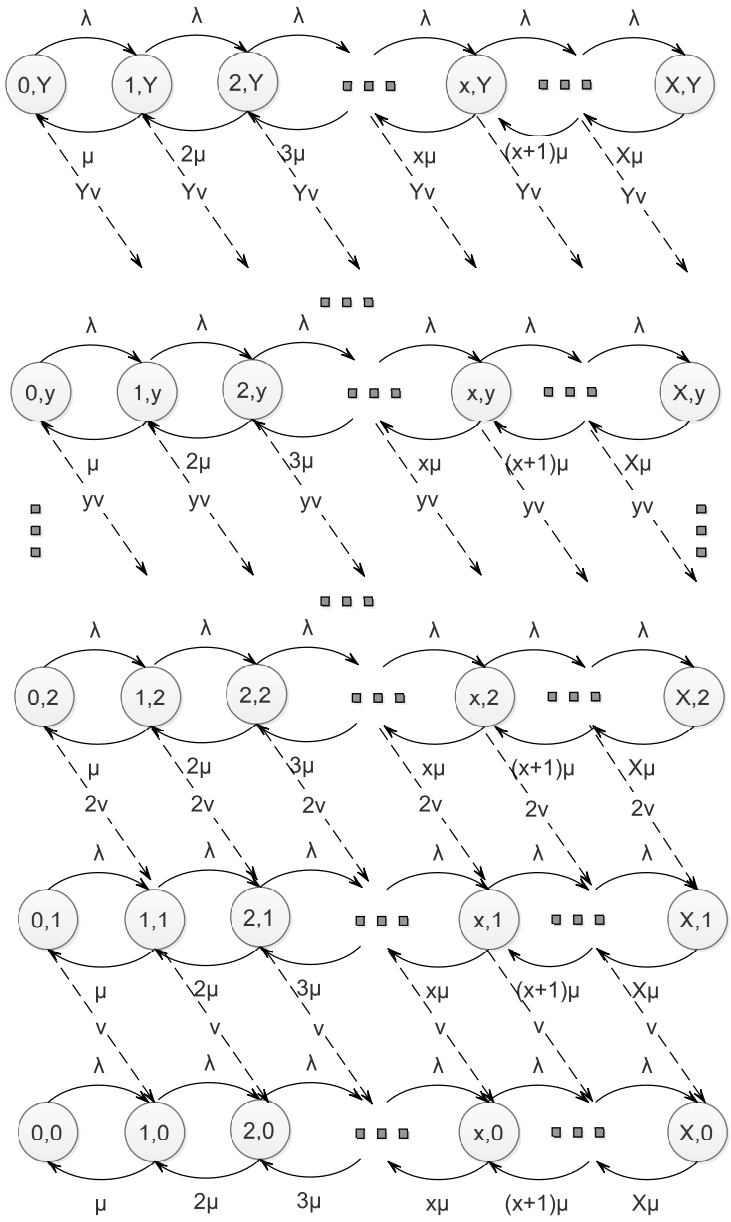


Figure 5.5: State transition diagram of considered system

transitions that refer to retrials of the flows that attempt to occupy the clusters. Thus, when a flow from the orbit queue manages to occupy a cluster, the orbit queue length reduces and the number of occupied clusters increases. The expected number of occupied clusters is equal to:

$$\mathbb{E}[X] = \sum_{(x,y)\in\mathbb{A}} x\pi(x,y), \tag{5.18}$$

whereas the expected length of the orbit queue is:

$$\mathbb{E}[Y] = \sum_{(x,y)\in\mathbb{A}} y\pi(x,y). \tag{5.19}$$

In order to derive the probabilities $\pi(x,y)$, we denote as $\boldsymbol{\pi}$ the steady state probability vector that can be ordered as:

$$\begin{gathered}\boldsymbol{\pi} = [\pi(0,0), \pi(0,1), \ldots, \pi(0,Y),\\ \vdots\\ \pi(1,0), \pi(1,1), \ldots, \pi(1,Y),\\ \vdots\\ \pi(X,0), \pi(X,1), \ldots, \pi(X,Y)],\end{gathered}$$

and solve the equation $\boldsymbol{\pi Q} = \mathbf{0}$ with the normalization condition $\boldsymbol{\pi 1} = 1$, where

$$\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{A}_{0,0} & \boldsymbol{A}_{0,1} & & & & & & \\ \boldsymbol{A}_{1,0} & \boldsymbol{A}_{1,1} & \boldsymbol{A}_{1,2} & & & & & \\ & \boldsymbol{A}_{2,1} & \boldsymbol{A}_{2,2} & \boldsymbol{A}_{2,3} & & & & \\ & & \ddots & \ddots & \ddots & & & \\ & & & \boldsymbol{A}_{y,y-1} & \boldsymbol{A}_{y,y} & \boldsymbol{A}_{y,y+1} & & \\ & & & & \ddots & \ddots & \ddots & \\ & & & & & \boldsymbol{A}_{Y,Y-1} & \boldsymbol{A}_{Y,Y} & \boldsymbol{A}_{Y,Y+1} \end{bmatrix}$$

is the generator matrix that consists of $\boldsymbol{A}_{y,y-1}, \boldsymbol{A}_{y,y}, \boldsymbol{A}_{y,y+1}$ matrices of order $(X+1)$ and $\mathbf{1} = [1, \ldots, 1]^T$ is the unit vector [123]. The values of the matrices are provided in [124]. Given the vector $\pi(x) = [\pi(x,0), \pi(x,1), \ldots, \pi(x,Y)]$, it holds that $\boldsymbol{\pi 1} = \sum_{x\in\mathbb{A}} \pi(x) = 1$.

5.5 Model validation and performance assessment

In this section, we validate the analytical models, investigate the performance of the MTFP algorithm in terms of blocking probability GoS, delay and energy efficiency considering different scenarios and demonstrate the convergence of the MTFP algorithm. We also investigate the tradeoff between the induced delay and the control overhead of the MTFP algorithm.

Table 5.2: Simulation parameters for model validation and performance evaluation of MTFP algorithm

Parameter	Value
LTE-A network settings	
N	2 MNOs
$\|\mathcal{N}\|$	2 eNBs
RBs per MNO	50 or 100 RBs
Bandwidth per MNO	10 MHz or 20 MHz
W	100 or 200 RBs
Modulation schemes	QPSK, 16-QAM, 64-QAM
Channel model	Rayleigh fading
Average SNR γ	10 dB
TTI	1 ms
OSP related settings	
$\|\mathcal{M}\|$	2 OSPs
Priority classes per OSP	2 ($p_j = 1$: high priority, $p_j = 2$: low priority)
Downlink data rates	0.5 Mb/s (low priority), 1 Mb/s (high priority)

5.5.1 Simulation setup

In all simulation scenarios, we consider a shared LTE-A network (Fig. 5.2) with $N = 2$ MNOs and $|\mathcal{M}| = 2$ OSPs that offer video streaming services, e.g, YouTube [125] or Skype [126]. Each OSP has $|\mathcal{K}| = 2$ priority classes that denote their users' subscription status, i.e., a high priority class with downlink data rate demand equal to 1 Mb/s, which includes premium users that require higher quality video, and a low priority class with 0.5 Mb/s, which refers to freemium users. High priority characterizes 50% of the flows, whereas the other 50% of the flows belong to the low priority class. For the high priority flows, we set the priority of the most preferred contracts as $p_j = 1$, whereas for the low priority flows, $p_j = 2$. In each VS allocation round, the value $v_n^{(m,j)}$ varies, as the number of RBs required to achieve the requested downlink data rate for a UE may vary, according to the downlink channel conditions that determine the MCS, as described in Section 5.2.2. Hence, the q values vary throughout the simulation.

The MNOs share their spectrum jointly operating $|\mathcal{N}| = 2$ eNBs. Each MNO contributes with 50 or 100 RBs, corresponding to bandwidth 10 MHz and 20 MHz, respectively [121]. A number W of 100 or 200 RBs is available in the shared spectrum pool. Furthermore, three modulation schemes are used, i.e., QPSK, 16-QAM and 64-QAM. Each modulation scheme is associated with a set of coding rates, defining an MSC determined by each UE according to the experienced SNR. Given a number of allocated RBs, the MCS determines the TBS, derived using the table provided in [106]. Using the TBS, the achievable downlink data rate is given by Eq.(1) with TTI equal to 1 ms. For the estimation of each UE's SNR, the Rayleigh fading channel model is used [127], with average SNR γ set to 10 dB [121]. The simulation parameters are summarized in Table 5.2.

In Section 5.5.2, we evaluate the proposed blocking probability GoS analysis provided in Section 5.4.1 and we assess the performance of MTFP algorithm in terms of GoS.

Considering the lack of resource allocation approaches for OSPs, we compare the MTFP algorithm with a best effort (BE) approach that allocates randomly the RBs to the flows without considering the OSPs' policies. In Section 5.5.7, we demonstrate the convergence of MTFP in a simple simulation scenario. In Section 5.5.4, motivated by the network neutrality issue that arises when multiple OSPs access a shared network, we examine the fairness in VS allocation with MTFP. In Section 5.5.5, we evaluate MTFP and BE in realistic scenarios, studying the network during a simulation period of two hours. Using various numbers of UEs, flow generation rates and VS allocation steps, we estimate the average delay induced when flows fail to receive resources in each VS allocation round and evaluate the analysis presented in Section 5.4.2. In Section 5.5.6, we study the performance of MTFP and BE in terms of energy efficiency. Finally, in Section 5.5.7, we study the tradeoff between the experienced delay and the control overhead in MTFP algorithm, estimating the control overhead for different effective throughput values in the RAN-VC paths.

5.5.2 GoS model validation and comparison with BE approach

We thereupon evaluate the GoS analysis assuming a shared network with $W = \{100, 200\}$ RBs and a number of $U = \{40, 60, 80, 100\}$ UEs (one flow corresponds to one UE). The flows are distinguished in two priority classes, as described in Section 5.5.1.

As it can be observed in Fig. 5.6, the simulation results corroborate our analysis. Moreover, in this figure, we may see that the MTFP algorithm outperforms the BE approach in all cases, achieving a GoS reduction of $23 - 38\%$, for $W = 100$, comparing the cases with $|\mathcal{J}| = 100$ and $|\mathcal{J}| = 40$, respectively. For $W = 200$, a reduction of $35 - 50\%$ is achieved. With MTFP, the exact number of RBs required to achieved the requested data rates in the eNBs that offer the best possible downlink channel conditions (enabling the use of higher MCS values) to the UEs are allocated. Furthermore, the GoS achieved by both approaches increases along with the number of the flows, as fewer flows can be served with the same number of RBs. Still, for the same W, the GoS of the MTFP algorithm is significantly lower than the GoS of the BE approach, as the available RBs are better utilized. It is also worth noting that, for high numbers of flows, i.e., higher than 60, the MTFP algorithm has better performance than the BE approach, even when the available resources are fewer.

We should also refer that the flows accommodated by the BE approach may belong to either of the two classes. Considering the case of $W = 200$ and $|\mathcal{J}| = 100$ flows, where GoS is equal to 0.43 and 0.67 for MTFP and BE, respectively, with MTFP, on average, 43 (i.e., 100·0.43) rejected flows belong to low priority class, whereas all high priority flows are accommodated, as each class has 50 flows and 57 (=100-43) flows receive RBs. In contrast, each of the 67 (i.e., 100·0.67) flows rejected when BE is applied may belong to either of the two classes.

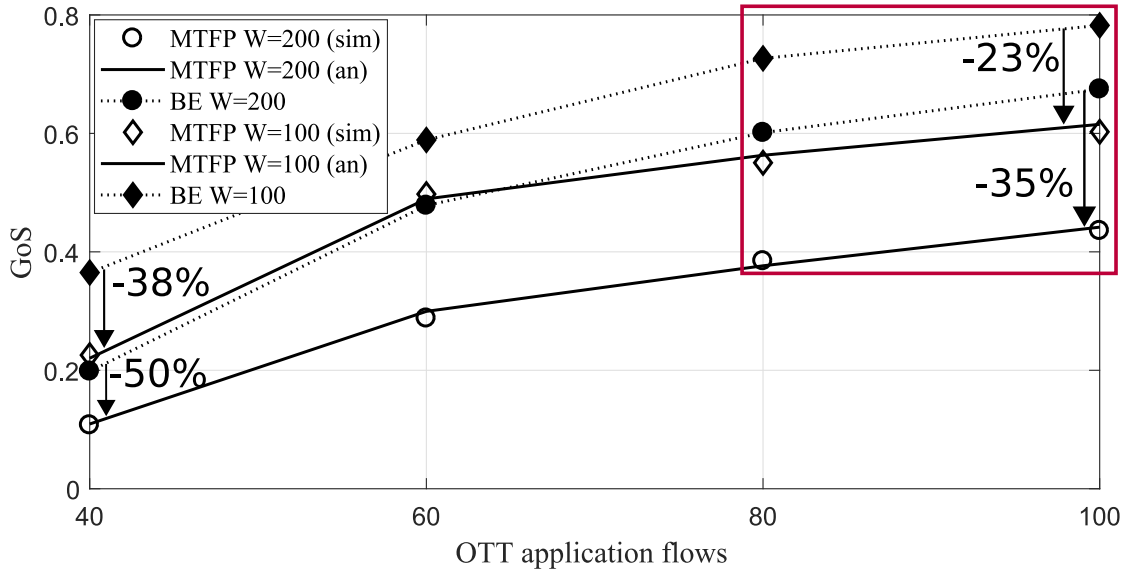


Figure 5.6: Grade-of-service vs. different numbers of OTT application flows

5.5.3 Study of convergence of MTFP algorithm

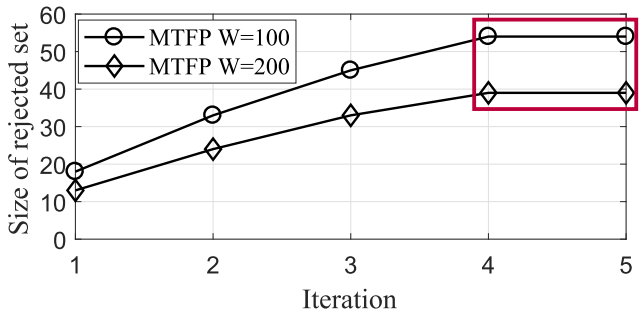
As stated in Proposition 1 (Section 5.3.3.1), the MTFP algorithm converges to a stable matching, when the size of the rejected set R_N stops increasing, i.e., the R_N has the same size in the last two iterations of the algorithm. We thereupon demonstrate the convergence of MTFP in a simple simulation scenario, where 40 flows request resources in a VS allocation round. Half of the flows of each priority class are new and request resources for the first time in this round. Each flow creates a preference list with $(|\mathcal{K}||\mathcal{N}|+1) = 5$ contracts, including the null contract.

In Fig. 5.7(a), we observe that the size of the R_N set increases from iterations 1 to 4, as there exist contracts submitted by the flows that are rejected by the eNBs. The flows that are rejected in an iteration submit the next most preferred contracts in the subsequent iteration. In the last iteration, the flows that submit contracts are accepted with the null contract, which denotes that all of the available RBs are already occupied, thus, they cannot be served with the requested data rates. As their contracts are accepted, the size of R_N remains the same in the last two iterations, showing the convergence to a solution that offers the minimum possible GoS, as depicted in Fig. 5.7(b).

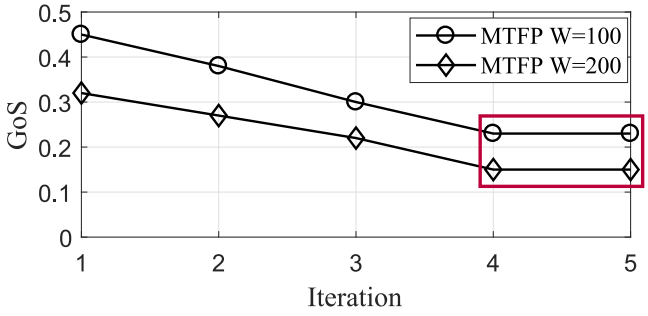
5.5.4 Study of fairness in VS allocation with MTFP algorithm

We next focus on the shared network scenario where $W = 100$ RBs are available. Aiming to assess the fairness levels in resource allocation when the MTFP algorithm is applied, we examine the GoS achieved for each OSP and MNO with respect to the number of OTT application flows in the network. In Fig. 5.8, the performance results of the MTFP algorithm in terms of fairness in GoS are demonstrated.

In Figs. 5.8(a) and 5.8(b), we see that MTFP achieves the same levels of GoS for all



(a) Size of rejected set R_N per iteration



(b) GoS per iteration

Figure 5.7: Convergence of the MTFP algorithm

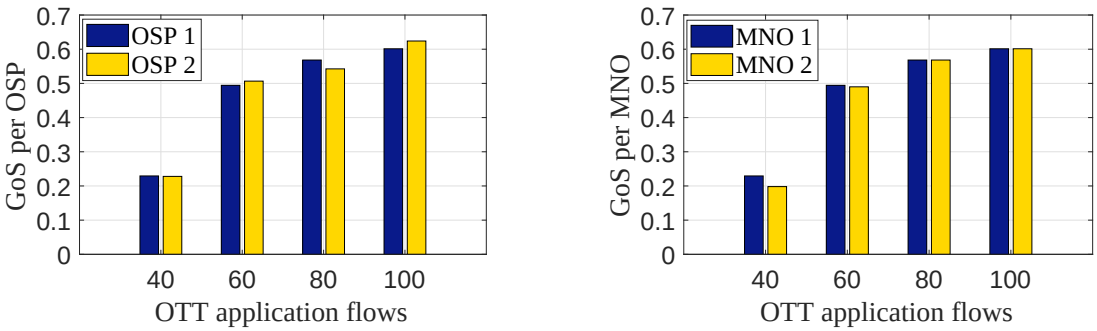
OSPs thus, the same number of each OSP's flows is served with the requested data rates. MTFP prioritizes the high priority flows in RB allocation but does not distinguish the different OSPs. Similarly, as each flow corresponds to a UE related to either of the two MNOs that share the network, MTFP does not prioritize the UEs of a specific MNO. For a quantitative measurement of the fairness level, we plot the fairness index θ of the GoS achieved for OSPs and MNOs, defined as [128]:

$$\theta = \frac{\left(\sum_{i=1}^{I} GoS_i\right)^2}{I \sum_{i=1}^{I} GoS_i^2}, \theta \in (0, 1], \tag{5.20}$$

where $I = |\mathcal{M}|$ for OSPs or $I = |\mathcal{N}|$ for MNOs. The highest fairness level is achieved when the θ value is equal to one for all OSPs or MNOs, whereas θ reduces when the GoS values are dispersed. As depicted in Fig. 5.8(c), MTFP results in similar GoS for all OSPs and MNOs in all cases, achieving θ values very close to 1 for both OSPs and MNOs.

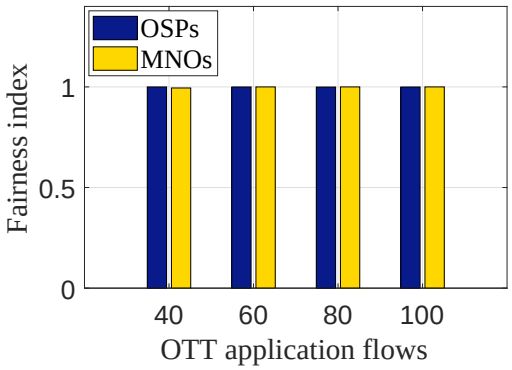
5.5.5 Delay model validation and study of induced delay

We next investigate the average delay experienced by the flows during a time period of two hours and evaluate the corresponding analytical model presented in Section 5.4.2. In total, $W = 100$ RBs are available in the considered shared network. A number of U UEs, out of which $U/2$ are related to each MNO, generate flows following a Poisson distribution with



(a) Grade-of-service per OSP

(b) Grade-of-service per MNO



(c) Fairness index θ

Figure 5.8: Fairness in GoS vs. number of OTT application flows

rate λ (flows/hour/UE). Each UE generates at least one flow for each OTT application, and, for a specific UE, flows of the same application have the same priority. The average number of high priority flows is equal to the average number of low priority flows generated in the simulation period, whereas half of the generated flows related to one OSP belong to high priority class. Each flow has an exponentially distributed duration with mean equal to $1/\mu = 180$ s. The mean value of the VS allocation step $\mathbb{E}[t]$ is set to 50 ms and 100 ms, providing a reasonable time frame for the information about the UEs to be transmitted to the VC, as determined by the CN congestion levels [129]. The value 100 ms can be considered as the upper bound for the delay in LTE-A networks [130].

We evaluate the delay analysis considering various values of the number of UEs U, OTT flow generation rates λ, and VS allocation step. As shown in Figs. 5.9 and 5.11, the analysis is verified by the match of theoretical and simulation results. We also study the effect of different numbers of UEs and OTT flow generation rates, comparing the MTFP algorithm with the BE approach.

5.5.5.1 Effect of different numbers of UEs

We study the effect of number of UEs that are connected to the considered shared LTE-A network on the delay experienced by the flows, using the MTFP and BE approaches. A number of $U = \{100, 200, \ldots, 500\}$ UEs and two different VS allocation steps, i.e., 50 and

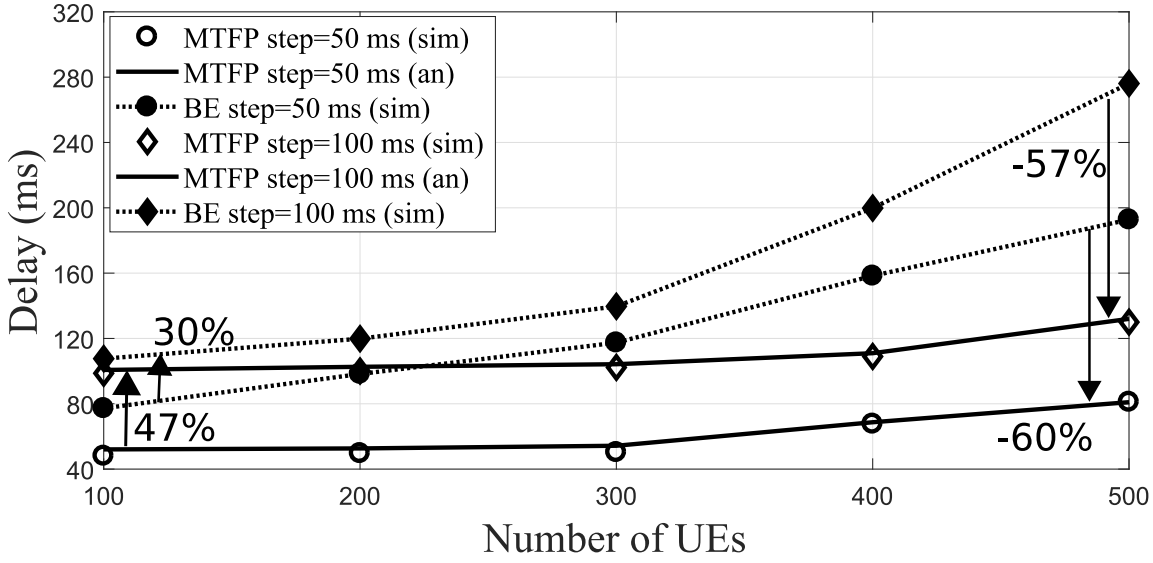


Figure 5.9: Delay vs. number of UEs

100 ms, are considered, simulating different CN congestion levels.

As illustrated in Fig. 5.9, the increase of the number of UEs leads to higher experienced delay, since more flows are generated and compete for resources. Still, MTFP achieves lower delay values than BE, reaching a reduction of up to 60% and 57% comparing to BE (for $U = 500$ and step values equal to 50 and 100 ms, respectively), as RBs are allocated in a way that the highest possible number of flows are accommodated in each VS allocation round. In contrast, the BE approach results in up to 137% and 112% higher delay for step values of 50 and 100 ms ($U = 500$), respectively, as it does not take into account the OSPs' performance goals and allocates randomly the RBs to the flows.

Moreover, for both schemes, the delay is higher when the step value increases, reaching values up to 47% and 30% higher for MTFP and BE ($U = 100$), respectively. As the information exchange takes longer to be completed, each round lasts longer and the impact of lost rounds on the experienced delay is higher, increasing the average delay experienced by the flows.

5.5.5.2 Effect of different OTT flow generation rates

We next focus on the effect of different flow generation rates on the experienced delay, using the MTFP and BE approaches. Assuming a number of $U = 500$ UEs, we set $\lambda = \{2, 4, 6, 8\}$ flows/hour per connected UE.

In Fig. 5.10, it can be observed that, for both approaches, the higher the number of flows generated by each UE, the higher the induced delay, as higher number of flows participate concurrently in VS allocation rounds, requesting resources in order to achieve the required data rates. As expected, the increase of step value affects the delay negatively. However, MTFP still achieves better performance, as it results in delay values 55%-60% and 32%-48% lower than those achieved by the BE approach for the step values of 50 ms

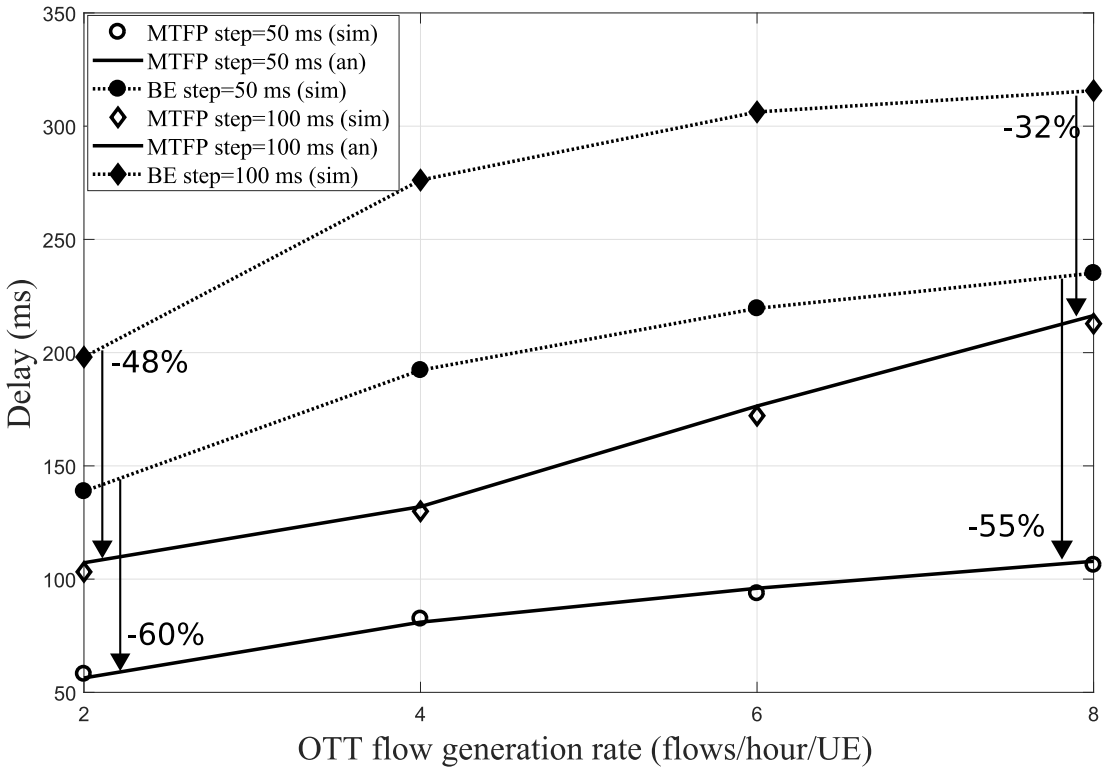


Figure 5.10: Delay vs. OTT flow generation rate

and 100 ms, respectively (BE results in delay values $121\% - 138\%$ and $48\% - 91\%$ higher than those of MTFP).

A closer inspection of the delay (Fig. 5.11) induced by the MTFP algorithm for the two different flow priority classes, i.e., high and low priority classes, shows that for the same step value, the delay experienced by high priority flows is lower than that of low priority flows, reaching a reduction of 35% and 37% for step values of 50 ms and 100 ms ($\lambda = 8$), respectively. This result corroborates that MTFP prioritizes the flows, allowing the high priority flows to receive resources more often throughout their duration. Still, the low priority flows manage to receive resources, though they experience higher delay.

Overall, it can be observed that the MTFP performance is affected by the CN and RAN congestion. The use of higher step values that correspond to longer transmission duration of flows' information and the co-existence of higher number of flows are two parameters that impact on GoS and delay. Even though MTFP manages to prioritize certain flows, it its still influenced by the end-to-end network congestion, stressing the need for VS allocation approaches that consider the OSPs' policies in resource allocation of both CN and RAN. Last, we should note that MTFP achieves flow prioritization without applying OSP prioritization, abiding by the network neutrality principle.

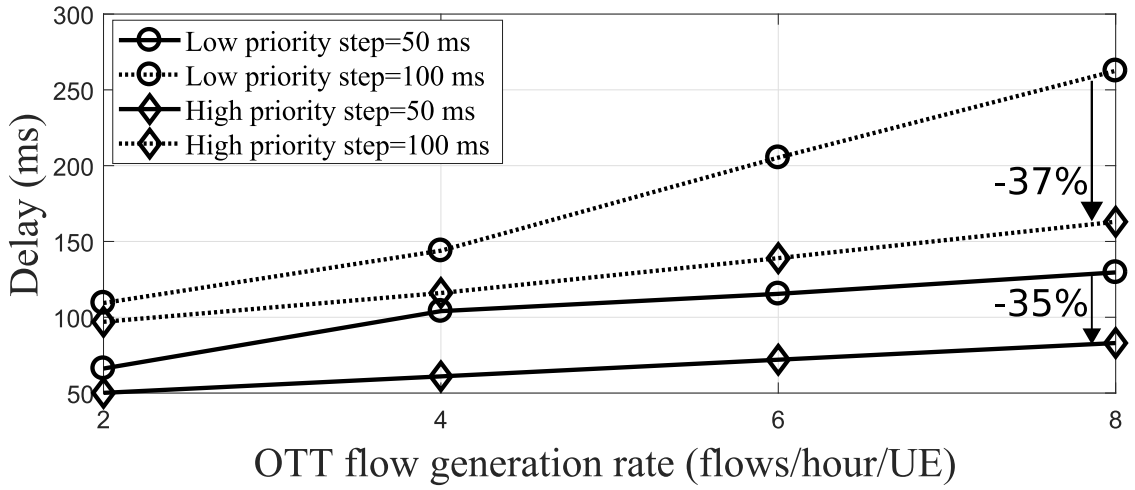


Figure 5.11: Delay vs. OTT flow generation rate per priority class using MTFP algorithm

5.5.6 Study of induced energy efficiency

We study the performance of the MTFP and BE approaches in terms of energy efficiency, considering different number of connected UEs (similarly as in the simulation scenario of Section 5.5.5.1) and different OTT flow generation rates (similarly as in the simulation scenario of Section 5.5.5.2). For the estimation of the energy efficiency, we set $P_C^{(n)} = 354.44$ W, $P_{Tx}^{(n)} = 46$ dBm, $\delta = 21.45$ and $a_n = 2\ \forall n \in \mathcal{N}$ [120].

In Fig. 5.12, we can observe that the MTFP algorithm outperforms the BE approach in terms of energy efficiency, reaching an increase of 93% and 96%, for step equal to 50 and 100 ms, respectively ($U = 100$). With MTFP, the RBs are allocated in accordance with the flows' downlink channel conditions and QoS demands and the total data rate increases, improving the energy efficiency. As the eNBs are always active and no switching off scheme is applied, i.e., P_n (Eq. (5.5)) is always considered, it is more efficient that more flows are served by each eNB n per VS allocation round.

We next focus on the effect of different flow generation rates. Fig. 5.13 demonstrates that the MTFP algorithm improves the energy efficiency by 74% and 76% ($\lambda = 8$) for step=50 and 100 ms, respectively, comparing to BE. Also, as λ increases, the energy efficiency improvement attenuates, as more RBs become occupied, providing the highest total data rate that is feasible per VS allocation round. When MTFP is applied and λ is higher than 4, it can be seen that although the higher step value (100 ms) produces higher delay (as shown in Fig. 5.10 presented in Section 5.5.5.2), it improves the energy efficiency up to 26% ($\lambda = 8$), as it leads to fewer rounds with low RB utilization.

5.5.7 Study of delay and control overhead tradeoff

As described in Section 5.3.3.2, in each VS allocation round, the MTFP algorithm requires the exchange of control messages. We next study the tradeoff between the experienced delay and the control overhead β in a shared network with $|\mathcal{N}| = 2$ eNBs, $U = 200$ UEs

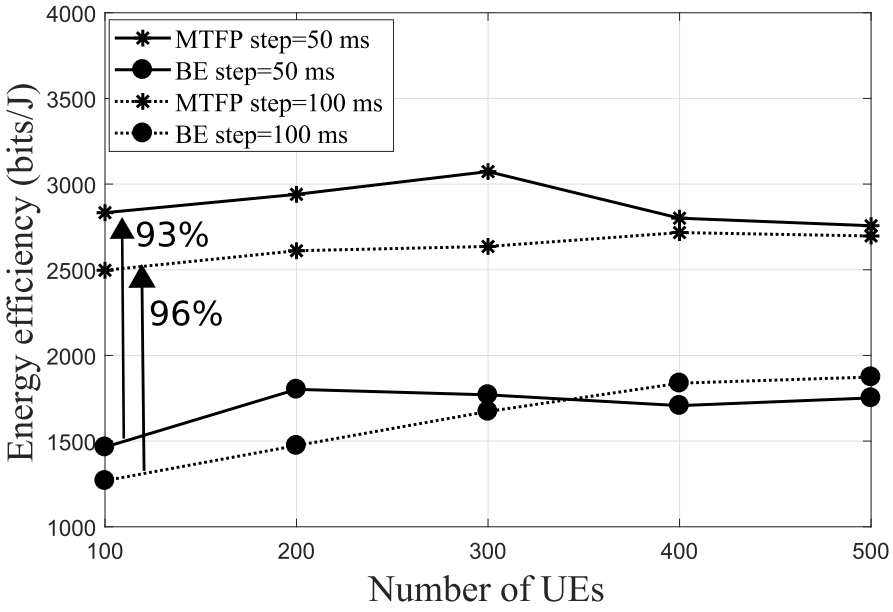


Figure 5.12: Energy efficiency vs. number of UEs

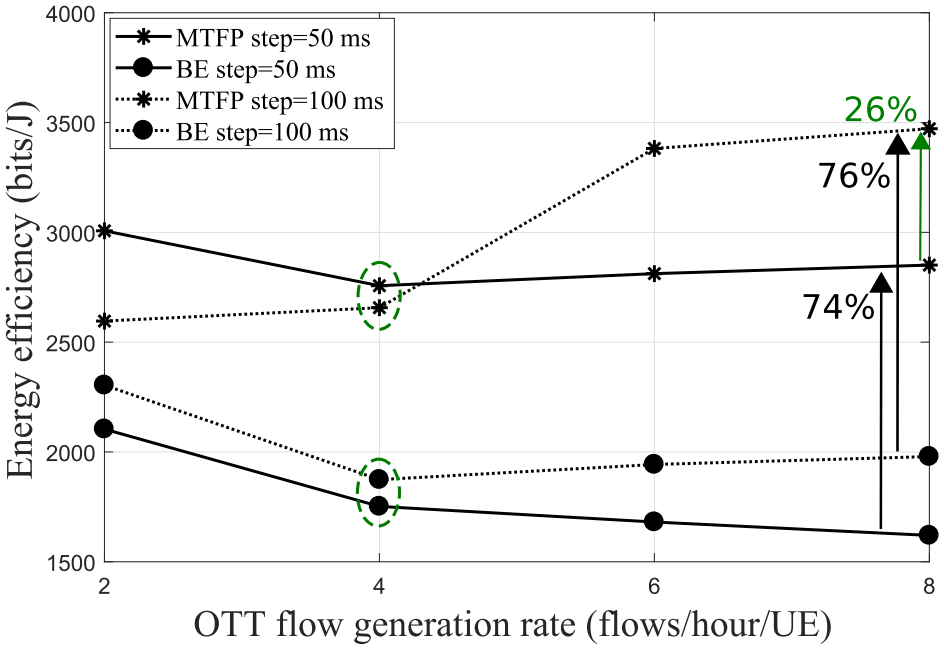


Figure 5.13: Energy efficiency vs. OTT flow generation rate

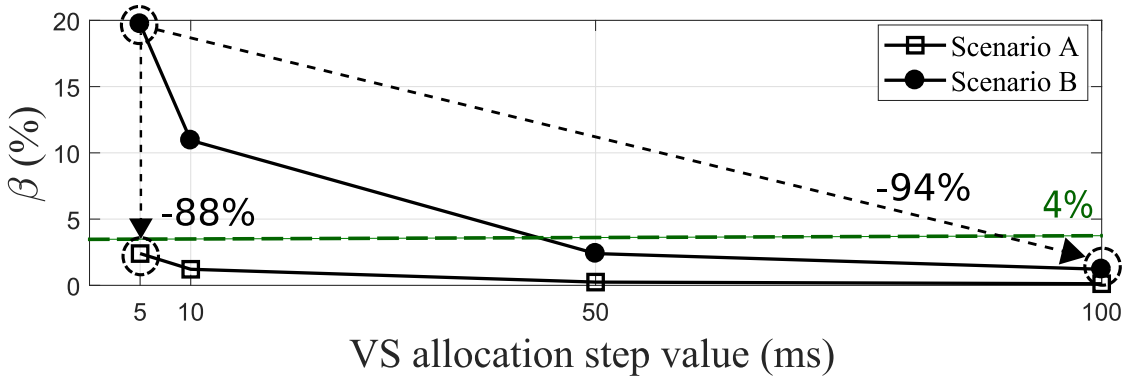


Figure 5.14: Control overhead vs. VS allocation step value

(one flow per UE) and a control packet size $l_{ctrl} = 256$ bytes. As UEs report CQIs to all eNBs and the VC reports to eNBs information about all UEs, in Eq. (5.4), we set $s_{ctrl} = U(|\mathcal{N}| + 1)l_{ctrl}$. Two scenarios with different r_e values are considered: i) scenario A, with $r_e = 10$ Gb/s, which may correspond to a network with a fiber link between eNBs and VC, and ii) scenario B, with $r_e = 1$ Gb/s, which may refer to a heterogeneous network, where the eNBs also communicate with small cells interconnected with wireless links and thus, multihop RAN-VC paths are created, whereas r_e is considered to be the minimum of the data rates at each hop.

Fig. 5.14 shows the β levels for both scenarios, assuming $l_{ctrl} = 256$ B and $\mathbb{E}[t] = \{5, 10, 50, 100\}$ ms. In the same figure, the threshold of 4%, which is an acceptable control overhead level for efficient bandwidth utilization [131] is also plotted. We can observe that β is lower in network A, where r_e is higher, reaching a reduction of 88% ($\mathbb{E}[t] = 5$ ms), comparing to network B, as more data packets are transmitted per round. Moreover, β reduces when higher step values are used, e.g., in network B, for $\mathbb{E}[t] = 100$ ms, it is up to 94% lower, comparing to $\mathbb{E}[t] = 5$ ms, as more data packets are sent with less frequent control message transmissions. Notably, although the increase of step values improves β, it induces higher delay for the flows (Section 5.5.5), showing a trade-off between reducing the experienced delay and restraining the overhead. Also, the existence of links with different data rates in multihop RAN-VC paths of heterogeneous networks impacts on the control overhead, which is higher than the threshold for small step values.

5.6 Chapter concluding remarks

In this chapter, the MTFP algorithm for OSP-oriented resource management in shared LTE-A networks and an analytical model for the induced GoS and experienced delay have been presented. We have extensively studied the performance of the proposed algorithm considering different network characteristics, i.e., different numbers of UEs generating OTT application flow, OSPs and MNOs, and different flow generation and VS allocation rates. The analytical and simulation results have shown that MTFP achieves better GoS,

delay and energy efficiency performance, compared to a best effort scheme. MTFP also prioritizes the flows according to the OSPs' policies, abiding by the network neutrality principle, i.e., it achieves similar GoS levels for all OSPs.

Although the MTFP performance deteriorates as the number of flows and the duration of VS allocation rounds, i.e., the network congestion, increase, MTFP manages to accommodate a higher number of flows than the best effort approach, achieving up to 50% lower GoS when 40 OTT application flows exist and 200 RBs are available. Furthermore, MTFP achieves up to 60% lower delay than the BE approach, when a VS allocation step size equal to 50 ms is used. For a step value equal to 100 ms, MTFP results in up to 96% higher energy efficiency comparing to the best effort approach. Moreover, it has been observed that with MTFP, in high data traffic cases, i.e., higher number of UEs or higher OTT flow generation rate per UE, the longer duration of VS allocation rounds may increase the delay but can improve the energy efficiency, achieving an increase of up to 26%.

As various stakeholders join the wireless market, offering innovative OTT services, and claim end-to-end resources over a shared network in order to serve their users, the network resource management scheme should respect both OSPs' policies and the network neutrality principle. Simultaneously, the user experience should not degrade, keeping the delay induced by the resource management scheme in acceptable levels. It should be also noted that the users' QoS demands should be accommodated without inflating the energy consumption of the mobile network, which is an important factor that affects the overall efficiency of a resource management scheme. To that end, the MTFP algorithm constitutes an efficient means to manage the VSs under the influence of the aforementioned multifaceted requirements.

Chapter 6

Conclusions and future work

In this chapter, the main contribution of the thesis is summarized and the derived conclusions are discussed (Section 6.1). Additionally, directions for future research on the topics studied in this thesis are presented (Section 6.2).

6.1 Thesis concluding remarks

In this thesis, we have presented a set of techniques that are meant to address the RRM issues that arise in modern mobile networks, as formed by the interactions between LTE-A and D2D technologies and between MNOs and OSPs as business stakeholders. Our study has elaborated on two research directions: i) the resource management in outband D2D communication that is shaped as a MAC design issue and ii) the resource management for OTT applications in multi-tenant mobile networks. The presented techniques can provide useful intuition towards the development of RRM solutions in 5G networks.

In the first part of the thesis that focuses on the outband D2D communication and is described in Chapters 3 and 4, two D2D MAC protocols, i.e., the ACNC-MAC and the SCD2D-MAC protocol, have been presented. They are based on the NC technique and leverage the use of relays in order to improved the throughput and the energy efficiency of the D2D cooperative network. The ACNC-MAC protocol has been also studied under the influence of the joint cellular-D2D system and the cellular network-related factors that affect the outband D2D performance. It has been proved to be beneficial in cases of high traffic conditions, whereas its performance is affected by the scheduling policy and the MCSs used for the downlink channel transmissions of the data that are subsequently exchanged between D2D pairs. The SCD2D-MAC protocol relies on the social ties created among the mobile users by virtue of the various social networking mobile applications in order to select the appropriate relays, promoting the use of friendly UEs as relays. The performance analysis of the two schemes, both theoretical and simulation-based, have resulted in the following observations:

(i) The D2D cooperation performance is affected by the number of the UEs that are selected to operate as relays. Although higher number of relays seems to improve the D2D throughput, the overall energy consumption may increase.

(ii) The integration of the social information of the UEs in the D2D MAC design is a beneficial complement to the relay selection strategy that prioritizes the relays that experience the best channel conditions and are capable to perform NC. The existence or lack of social ties between the D2D pairs and their relays affects the energy consumption levels during cooperation.

(iii) When downlink transmission of data to the UEs occurs concurrently with the D2D data exchange, the UEs with better downlink channel conditions, i.e., higher MCSc, experience higher throughput with the MT scheduler, whereas for UEs with poor downlink channel conditions, the PF scheduler is preferable, for all traffic load levels.

In the second part of the thesis, elaborated in Chapter 5, we focus on the QoS provision for the OTT application users in shared mobile networks via the allocation of proper VSs. The MTFP algorithm is presented and its performance is evaluated by means of theoretical models and simulations. This algorithm enables the intervention of the OSP in the VS allocation without violating the network neutrality principle. Our study has enabled us to derive the following conclusions regarding the OSP-oriented resource management:

(i) The comparison of the MTFP algorithm with the best-effort approach has demonstrated that MTFP effectively considers the flows' priorities and is able to apply the OSPs' policies while also guaranteeing an equal treatment of all OSPs. It manages to accommodate a higher number of ows, improving the GoS, the experienced delay and the network energy efficiency levels, even in high load scenarios.

(ii) The framework of matching theory in the particular resource management problem enables the intervention of OSPs with relatively low overhead, which is related to the network capabilities, i.e., the achievable data rates in the paths from the RAN to the VC where the control messages circulate, and the network congestion levels, i.e., the number of OTT application flows. It has been observed that the use of higher step values, i.e., when the VS allocation rounds are performed with lower frequency, results in lower overhead. However, it increases the experienced delay, thus, proper selection of the step value is required in order to balance this trade-off.

6.2 Directions for future work

As the previous chapters were devoted to the detailed description of the technical aspects of our contributions, in order to round up the presentation of our study, we thereupon discuss several interesting directions for future research on the issues that our work has not yet covered.

(i) Regarding the resource management in D2D communication, our study has focused on the D2D data exchange over unlicensed spectrum inside an LTE-A network, whereas the proposed MAC schemes are compliant with the IEEE 802.11 specification (Wi-Fi). However, there also exist other technologies for outband access, such

as 802.15.1 (Bluetooth) and 802.15.4 (ZigBee), which are used in 2.4GHz industrial, scientific and medical (ISM) bands and 5GHz unlicensed national information infrastructure (U-NII) bands. Moreover, following the development of LTE-A technologies, such as the carrier aggregation that allows the MNOs to combine a number of separate LTE carriers, novel proposals for the operation of LTE in unlicensed spectrum (LTE-U) have appeared. Qualcomm has first proposed the utilization of the 5 GHz band employed by IEEE 802.11ac compliant Wi-Fi equipment in order to increase network coverage and capacity [132]. Also, 3GPP has standardized the licensed assisted access (LAA) and the operation of LTE in the Wi-Fi bands using the listen-before-talk (LBT) contention based protocol (LTE Release 13 [133]).

The LTE-U enables the users to access both licensed and unlicensed spectrum under a unified LTE network infrastructure, whereas LBT is designed to coexist with other Wi-Fi devices on the same band. In a D2D network where LTE-U and Wi-Fi enabled UEs coexist, the MAC solutions require time synchronization between Wi-Fi and LTE. Hence, the use of other D2D MAC schemes such as those proposed in this thesis would require proper adaptation to the LTE-U mechanisms. The LTE-U transmissions can dynamically avoid overlapping with Wi-Fi transmissions if an adequate number of frequencies are available. In case that no channel is available, the LTE-U transmission can be adapted in a way that the channel is shared fairly with Wi-Fi via the carrier sense adaptive transmission method. With either method, there should be a coordination, possibly performed by the eNB, in order to determine the frequencies and the time slots allocated to each type of transmission.

As the joint LTE-U/Wi-Fi transmission coordination requires the intervention of the eNB or a centralized controller, additional overhead might be induced due to the information exchange between the coordinator and the UEs. The channel signaling might affect the performance of the D2D MAC protocol that targets the scenario of coexisting LTE-U and Wi-Fi transmissions. Therefore, the D2D MAC design should be studied in the joint LTE-U/LTE-A context.

(ii) Regarding the resource management the OSPs elaborated in the second part of the thesis, it should be highlighted that it is an issue that has arised recently. In this thesis, we have focused on the allocation of VSs in a shared network with a single layer of cells. However, the network densification and the deployment of heterogeneous infrastructure, with the addition of small cells and Wi-Fi APs in order to extend the cellular network coverage and capacity, are expected to culminate in the next few years [134]. As the OTT applications become more and more pervasive, the fair sharing of network resources between OSPs can be challenging due to the requirement for joint application of user policies determined by different OSPs over heterogeneous infrastructure.

At this point, it should be noted that the proposed matching theoretic method can be adapted to the scenario of heterogeneous mobile networks. Still, the additional

overhead in each VS allocation round and the delay induced by the application of the method should be studied under the influence of the new dense infrastructure. Furthermore, the preferences of the OSPs can be more complicated as different types of network elements with different types of resources may co-exist, e.g., RBs in small cells or time slots in APs. Thus, the proposed method may have to be refined in order to address the OSPs' requirements over a heterogeneous network.

Last, the network neutrality issue that has been under discussion for several years now can be considered as an additional challenge. We should mention that lately, the federal communications commission (FCC) has repealed the network neutrality rules imposed to ISPs, opening the road to paid prioritization [135]. However, the network neutrality remains an issue and the strategies that will be adopted by the ISPs/MNOs are not straightforward, as it is still under study if paid prioritization is overall beneficial for the end users.

To conclude, it should be noted that the concerns that are mentioned so far are an indicative subset of the challenges that the new technologies and the entry of new stakeholders pose in the upcoming generation of mobile networks. The resource management methods we have proposed in this thesis cannot claim to be absolute and unique solutions to the problems under study. Nevertheless, we believe that the presented methods can be a valuable contribution to the improvement of 5G mobile networks and that our study can provide insights towards the design of efficient resource management techniques that leverage the capabilities of the future mobile networks and are able to ensure the provision of high-quality services to the end users.

References

[1] CISCO. Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2015–2020. 2016.

[2] S. Esselaar, S. Song, and C. Stork. Freemium Internet: Next Generation Business Model to connect next billion. In *28th European Regional Conference of the International Telecommunications Society (ITS): "Competition and Regulation in the Information Age"*, 2017.

[3] LTE OFDM Technology. `https://www.tutorialspoint.com/lte/lte\_ofdm\_technology.htm`, 2018. Accessed on: 2018-07-01.

[4] E. Dahlman, S. Parkvall, and J. Skold. *4G, LTE-Advanced Pro and the Road to 5G*. Academic Press, 2016.

[5] J. G. Andrews, S. Buzzi, W. Choi, S. V Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang. What will 5G be? *IEEE Journal on Selected Areas in Communications*, 32(6):1065–1082, 2014.

[6] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski. Five disruptive technology directions for 5G. *IEEE Communications Magazine*, 52(2):74–80, Feb. 2014.

[7] A. Asadi, Q. Wang, and V. Mancuso. A survey on Device-to-Device communication in cellular networks. *IEEE Communications Surveys & Tutorials*, 16(4):1801–1819, Fourthquarter 2014.

[8] T. Wang, Y. Sun, L. Song, and Z. Han. Social Data Offloading in D2D-Enhanced Cellular Networks by Network Formation Games. *IEEE Transactions on Wireless Communications*, 14(12):7004–7015, Dec. 2015.

[9] 3rd Generation Partnership Project. LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Frequency (RF) system scenarios (3GPP TR 36.942 version 12.0.0 Release 12). Oct. 2014.

[10] A. Khan, W. Kellerer, K. Kozu, and M. Yabusaki. Network sharing in the next mobile network: TCO reduction, management flexibility, and operational independence. *IEEE Communications Magazine*, 49(10):134–142, Oct. 2011.

[11] C. Liang, F. R. Yu, and X. Zhang. Information-centric network function virtualization over 5G mobile wireless networks. *IEEE Network*, 29(3):68–74, 2015.

[12] W. Green, B. Lancaster, and J. Sladek. Over the Top Services. *Pipeline Magazine*, 12, 2007.

[13] S. Greenstein, M. Peitz, and T. Valletti. Net neutrality: A fast lane to understanding the trade-offs. *Journal of Economic Perspectives*, 30(2):127–50, 2016.

[14] S. Sesia, M. Baker, and I. Toufik. *LTE-the UMTS long term evolution: from theory to practice*. John Wiley & Sons, 2011.

[15] The 3rd Generation Partnership Project (3GPP). Evolved universal terrestrial radio access (E-UTRA); long term evolution (LTE) physical layer; general description. *TS 36.201*, March 2009.

[16] The 3rd Generation Partnership Project (3GPP). LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) radio transmission and reception, version 14.5.0, release 14. *TS 136.101*, Nov. 2017.

[17] F. Capozzi, G. Piro, L. A. Grieco, G. Boggia, and P. Camarda. Downlink Packet Scheduling in LTE Cellular Networks: Key Design Issues and a Survey. *IEEE Communications Surveys & Tutorials*, 15(2):678–700, Second 2013.

[18] 3rd Generation Partnership Project. Technical Specification Group Radio Access Network; Study on LTE Device to Device Proximity Services; Radio Aspects (3GPP TR 36.843 version 12.0.1 Release 12). March 2014.

[19] J. Liu, N. Kato, J. Ma, and N. Kadowaki. Device-to-Device Communication in LTE-Advanced Networks: A Survey. *IEEE Communications Surveys & Tutorials*, 17(4):1923–1940, 2014.

[20] NOKIA White Paper: Optimising Spectrum Utilisation towards 2020. `http://networks.nokia.com/file/30301/optimising-spectrum-utilisation-towards-2020`, 2014. Accessed on: 2018-07-01.

[21] D. Camps-Mur, A. Garcia-Saavedra, and P. Serrano. Device-to-Device Communications with Wi-Fi Direct: Overview and Experimentation. *IEEE Wireless Communications*, 20(3):96–104, June 2013.

[22] J. Qiao, X. Shen, J. Mark, Q. Shen, Y. He, and L. Lei. Enabling Device-to-Device Communications in Millimeter-Wave 5G Cellular Networks. *IEEE Communications Magazine*, 53(1):209–215, Jan. 2015.

[23] C. Xu, L. Song, and Z. Han. *Resource Management for Device-to-Device Underlay Communication*. Springer, 2014.

[24] A. Antonopoulos, E. Kartsakli, and C. Verikoukis. Game Theoretic D2D Content Dissemination in 4G Cellular Networks. *IEEE Communications Magazine*, 52(6):125–132, June 2014.

[25] X. Lin, J. G. Andrews, and A. Ghosh. Spectrum sharing for device-to-device communication in cellular networks. *IEEE Transactions on Wireless Communications*, 13(12):6727–6740, 2014.

[26] A. Asadi, Q. Wang, and V. Mancuso. A Survey on Device-to-Device Communication in Cellular Networks. *IEEE Communications Surveys & Tutorials*, 16(4):1801–1819, Fourthquarter 2014.

[27] IEEE Standard for Inf. technology–Telecommun. and Inf. Exchange between systems, Local and Metropolitan Area Networks–Specific requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications. *IEEE Std 802.11-2012 (Revision of IEEE Std 802.11-2007)*, pages 1–2793, March 2012.

[28] T. Wang, Y. Sun, L. Song, and Z. Han. Social Data Offloading in D2D-Enhanced Cellular Networks by Network Formation Games. *IEEE Transactions on Wireless Communications*, 14(12):7004–7015, Dec. 2015.

[29] N. Golrezaei, P. Mansourifard, A.F. Molisch, and A.G. Dimakis. Base-Station Assisted Device-to-Device Communications for High-Throughput Wireless Video Networks. *IEEE Transactions on Wireless Communications*, 13(7):3665–3676, July 2014.

[30] M. N. Tehrani, M. Uysal, and H. Yanikomeroglu. Device-to-Device Communication in 5G Cellular Networks: Challenges, Solutions, and Future Directions. *IEEE Communications Magazine*, 52(5):86–92, May 2014.

[31] Y. Zhang, L. Song, W. Saad, Z. Dawy, and Z. Han. Contract-Based Incentive Mechanisms for Device-to-Device Communications in Cellular Networks. *IEEE Journal on Selected Areas in Communications*, 33(10):2144–2155, Oct. 2015.

[32] A. Asadi, V. Sciancalepore, and V. Mancuso. On the Efficient Utilization of Radio Resources in Extremely Dense Wireless Networks. *IEEE Communications Magazine*, 53(1):126–132, Jan. 2015.

[33] A. Asadi and V. Mancuso. On the Compound Impact of Opportunistic Scheduling and D2D Communications in Cellular Networks. In *Proc. ACM International Conference on Modeling, Analysis & Simulation of Wireless and Mobile Systems*, pages 279–288, 2013.

[34] S. Andreev, O. Galinina, A. Pyattaev, K. Johnsson, and Y. Koucheryavy. Analyzing Assisted Offloading of Cellular User Sessions onto D2D Links in Unlicensed Bands. *IEEE Journal on Selected Areas in Communications*, 33(1):67–80, Jan. 2015.

[35] R. Ahlswede, C. Ning, S.-Y.R. Li, and R.W. Yeung. Network Information Flow. *IEEE Transactions on Information Theory*, 46(4):1204–1216, July 2000.

[36] A. Pyattaev, O. Galinina, S. Andreev, M. Katz, and Y. Koucheryavy. Understanding Practical Limitations of Network Coding for Assisted Proximate Communication. *IEEE Journal on Selected Areas in Communications*, 33(2):156–170, Feb. 2015.

[37] M. Sami, N. K. Noordin, M. Khabazian, F. Hashim, and S. Subramaniam. A Survey and Taxonomy on Medium Access Control Strategies for Cooperative Communication in Wireless Networks: Research Issues and Challenges. *IEEE Communications Surveys & Tutorials*, 18(4):2493–2521, Fourthquarter 2016.

[38] S. Katti, H. Rahul, W. Hu, D. Katabi, M. Médard, and J. Crowcroft. XORs in the air: Practical wireless network coding. In *ACM SIGCOMM computer communication review*, volume 36, pages 243–254. ACM, 2006.

[39] J. Zhang, Y. P. Chen, and I. Marsic. MAC-layer Proactive Mixing for Network Coding in Multi-hop Wireless Networks. *Computer Networks*, 54(2):196 – 207, 2010.

[40] A. Antonopoulos, C. Verikoukis, C. Skianis, and Ö. B. Akan. Energy Efficient Network Coding-based MAC for Cooperative ARQ Wireless Networks. *Ad Hoc Networks*, 11(1):190 – 200, 2013.

[41] X. Wang and J. Li. Network Coding Aware Cooperative MAC Protocol for Wireless Ad Hoc Networks. *IEEE Transactions on Parallel and Distributed Systems*, pages 167–179, Jan. 2014.

[42] S. Zhang, S. C. Liew, and P. P. Lam. Hot topic: Physical-layer network coding. In *Proceedings of the 12th annual international conference on Mobile computing and networking*, pages 358–365. ACM, 2006.

[43] S. Wang, Q. Song, X. Wang, and A. Jamalipour. Distributed MAC Protocol Supporting Physical-Layer Network Coding. *IEEE Transactions on Mobile Computing*, 12(5):1023–1036, May 2013.

[44] A. Munari, F. Rossetto, and M. Zorzi. Phoenix: making cooperation more efficient through network coding in wireless networks. *IEEE Transactions on Wireless Communications*, 8(10), 2009.

[45] Ericsson. Ericsson Mobility Report–MWC Edition. `http://www.ericsson.com/ericsson-mobility-report`, 2015. Accessed on: 2018-07-01.

[46] Z. Wang, X. Zhou, D. Zhang, Z. Yu, and Daqiang Zhang. SOCKER: Enhancing Face-to-Face Social Interaction Based on Community Creation in Opportunistic Mobile Social Networks. *Wireless Personal Communications*, 78(1):755–783, Sep. 2014.

[47] E. Z. Tragos, V. Angelakis, A. Fragkiadakis, D. Gundlegard, C.-S. Nechifor, G. Oikonomou, H. C. Pohls, and A. Gavras. Enabling reliable and secure IoT-based

smart city applications. In *IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, pages 111–116, 2014.

[48] R.-S. Cheng, C.-M. Huang, and G.-S. Cheng. A D2D Cooperative Relay Scheme for Machine-to-Machine Communication in the LTE-A Cellular Network. In *IEEE International Conference on Information Networking (ICOIN)*, pages 153–158, 2015.

[49] E. Z. Tragos and V. Angelakis. Cognitive Radio Inspired M2M Communications. In *IEEE 16th International Symposium on Wireless Personal Multimedia Communications (WPMC)*, pages 1–5, 2013.

[50] L. Lei, Z. Zhong, C. Lin, and X. Shen. Operator Controlled Device-to-Device Communications in LTE-Advanced Networks. *IEEE Wireless Communications*, 19(3):96–104, June 2012.

[51] B. Zhang, Y. Li, D. Jin, P. Hui, and Z. Han. Social-Aware Peer Discovery for D2D Communications Underlaying Cellular Networks. *IEEE Transactions on Wireless Communications*, 14(5):2426–2439, May 2015.

[52] Y. Cao, T. Jiang, X. Chen, and J. Zhang. Social-Aware Video Multicast Based on Device-to-Device Communications. *IEEE Transactions on Mobile Computing*, PP(99):1–1, July 2015.

[53] Y. Zhao, Y. Li, H. Mao, and N. Ge. Social Community Aware Long-Range Link Establishment for Multi-hop D2D Communication Networks. In *IEEE International Conference on Communications (ICC)*, pages 2961–2966, June 2015.

[54] Y. Cai, D. Wu, and L. Zhou. Green Resource Sharing for Mobile Device-to-Device Communications. In *IEEE Wireless Communications and Networking Conference (WCNC)*, pages 3142–3147, Apr. 2014.

[55] Y. Li, S. Su, and S. Chen. Social-Aware Resource Allocation for Device-to-Device Communications Underlaying Cellular Networks. *IEEE Wireless Communications Letters*, 4(3):293–296, June 2015.

[56] O. Semiari, W. Saad, S. Valentin, M. Bennis, and H.V. Poor. Context-Aware Small Cell Networks: How Social Metrics Improve Wireless Resource Allocation. *IEEE Transactions on Wireless Communications*, 14(11):5927–5940, Nov 2015.

[57] L. Wang, H. Wu, W. Wang, and K.-C. Chen. Socially Enabled Wireless Networks: Resource Allocation via Bipartite Graph Matching. *IEEE Communications Magazine*, 53(10):128–135, 2015.

[58] A. Antonopoulos, E. Kartsakli, and C. Verikoukis. Game Theoretic D2D Content Dissemination in 4G Cellular Networks. *IEEE Communications Magazine*, 52(6):125–132, June 2014.

[59] Y. Jiang, Q. Liu, F. Zheng, X. Gao, and X. You. Energy Efficient Joint Resource Allocation and Power Control for D2D Communications. *IEEE Transactions on Vehicular Technology*, PP(99):1–1, 2015.

[60] M. Zhang, X. Chen, and J. Zhang. Social-aware Relay Selection for Cooperative Networking: An Optimal Stopping Approach. In *IEEE International Conference on Communications (ICC)*, pages 2257–2262, June 2014.

[61] X. Chen, B. Proulx, X. Gong, and J. Zhang. Exploiting Social Ties for Cooperative D2D Communications: A Mobile Social Networking Case. *IEEE/ACM Transactions on Networking*, 23(5):1471–1484, Oct. 2015.

[62] T. Han and N. Ansari. Heuristic Relay Assignments for Green Relay Assisted Device to Device Communications. In *IEEE Global Communications Conference (GLOBECOM)*, pages 468–473, Dec. 2013.

[63] Body of European Regulators for Electronic Communications. BEREC-RSPG report on infrastructure and spectrum sharing in mobile/wireless networks". *BEREC*, June 2011.

[64] OECD Publishing. Wireless Market Structures and Network Sharing. *OECD Digital Economy Papers*, (243), 2014.

[65] 3rd Generation Partnership Project. 3GPP; Technical Specification Group Services and System Aspects; Network Sharing; Architecture and Functional Description; (Release 14). March 2017.

[66] Z. Feng, C. Qiu, Z. Feng, Z. Wei, W. Li, and P. Zhang. An Effective Approach to 5G: Wireless Network Virtualization. *IEEE Communications Magazine*, 53(12):53–59, Dec. 2015.

[67] X. Zhou, R. Li, T. Chen, and H. Zhang. Network slicing as a service: enabling enterprises' own software-defined cellular networks. *IEEE Communications Magazine*, 54(7):146–153, 2016.

[68] A. Gudipati, D. Perry, Li E. Li, and S. Katti. SoftRAN: Software Defined Radio Access Network. In *Proceedings of the Second ACM SIGCOMM Workshop on Hot Topics in Software Defined Networking*, pages 25–30, Aug. 2013.

[69] X. Foukas, M. K. Marina, and K. Kontovasilis. Orion: RAN Slicing for a flexible and cost-effective multi-service mobile network architecture. In *Proc. of Annual Int. Conf. on Mobile Computing and Networking*, pages 127–140. ACM, 2017.

[70] M. Yang, Y. Li, D. Jin, L. Su, S. Ma, and L. Zeng. OpenRAN: a software-defined ran architecture via virtualization. *ACM SIGCOMM Computer Communication Review*, 43(4):549–550, 2013.

[71] I. F. Akyildiz, P. Wang, and S.-C. Lin. SoftAir: A software defined networking architecture for 5G wireless systems. *Computer Networks*, 85:1–18, 2015.

[72] N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, S. Shenker, and J. Turner. OpenFlow: enabling innovation in campus networks. *ACM SIGCOMM Computer Communication Review*, 38(2):69–74, 2008.

[73] V. G. Nguyen, A. Brunstrom, K. J. Grinnemo, and J. Taheri. SDN/NFV-based Mobile Packet Core Network Architectures: A Survey. *IEEE Communications Surveys & Tutorials*, PP(99):1–1, 2017.

[74] Business of Apps article: Facebook leads list of top smartphone apps in 2016, but google dominates number of apps in ranking. `http://www.businessofapps.com/facebook-leads-list-of-top-smartphone-apps-in-2016-but-google-dominates-number-of-apps-in-ranking/`, 2017. Accessed on: 2018-07-01.

[75] C.H. Liu and K.C. Huang. The Hard Decision of Mobile Operators: A Dumb Pipe or a Value-Added Service Provider. `https://www.econstor.eu/bitstream/10419/168492/1/Huang-Liu.pdf`, 2017. Accessed on: 2018-07-01.

[76] M. Vincenzi, A. Antonopoulos, E. Kartsakli, J. Vardakas, L. Alonso, and C. Verikoukis. Multi-tenant Slicing for Spectrum Management on the Road to 5G. *IEEE Wireless Communications*, 24(5):118–125, 2017.

[77] T. Janevski. *Internet Technologies for Fixed and Mobile Networks*. Artech House, 2015.

[78] Nicholas Economides and Joacim Tåg. Network neutrality on the internet: A two-sided market analysis. *Information Economics and Policy*, 24(2):91–104, 2012.

[79] R. T. B. Ma, J. Wang, and D. M. Chiu. Paid Prioritization and its impact on Net Neutrality. *IEEE Journal on Selected Areas in Communications*, 35(2):367–379, Feb. 2017.

[80] P. Trakas, F. Adelantado, N. Zorba, and C. Verikoukis. A Quality of Experience-aware Association Algorithm for 5G heterogeneous networks. In *IEEE International Conference on Communications (ICC)*, pages 1–6, May 2017.

[81] A. Ahmad, A. Floris, and L. Atzori. QoE-centric Service Delivery: A Collaborative Approach among OTTs and ISPs. *Computer Networks*, 110:168–179, 2016.

[82] A. Antonopoulos, E. Kartsakli, C. Perillo, and C. Verikoukis. Shedding Light on the Internet: Stakeholders and Network Neutrality. *IEEE Communications Magazine*, 55(7):216–223, May 2017.

[83] P. Di Francesco, J. Kibiłda, F. Malandrino, N. J. Kaminski, and L. A. DaSilva. Sensitivity Analysis on Service-Driven Network Planning. *IEEE/ACM Transactions on Networking*, 25(3):1417–1430, June 2017.

[84] 3rd Generation Partnership Project. 3GPP; Technical Specification Group Services and System Aspects; System Architecture for the 5G System; Stage 2 (Release 15). June 2017.

[85] H. H. Gharakheili, A. Vishwanath, and V. Sivarama. Perspectives on Net Neutrality and Internet Fast-Lanes. *ACM Computer Communications Review*, 46(1):64–69, Jan. 2016.

[86] P. K. Agyapong, M. Iwamura, D. Staehle, W. Kiess, and A. Benjebbour. Design Considerations for a 5G Network Architecture. *IEEE Communications Magazine*, 52(11):65–75, 2014.

[87] Y. C. Wang and T. Y. Tsai. A Pricing-Aware Resource Scheduling Framework for LTE Networks. *IEEE/ACM Transactions on Networking*, 25(3):1445–1458, June 2017.

[88] M. Srinivasan, V. J. Kotagi, and C. S. R. Murthy. A Q-Learning Framework for User QoE Enhanced Self-Organizing Spectrally Efficient Network Using a Novel Inter-Operator Proximal Spectrum Sharing. *IEEE Journal on Selected Areas in Communications*, 34(11):2887–2901, Nov. 2016.

[89] M. Kalil, A. Shami, and Y. Ye. Wireless Resources Virtualization in LTE Systems. In *IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 363–368, 2014.

[90] Y. Xiao, Z. Han, C. Yuen, and L. A. DaSilva. Carrier Aggregation Between Operators in Next Generation Cellular Networks: A Stable Roommate Market. *IEEE Transactions on Wireless Communications*, 15(1):633–650, Jan. 2016.

[91] F. Fu and U.C. Kozat. Wireless Network Virtualization as A Sequential Auction Game. In *Proc. IEEE INFOCOM*, pages 1–9, Mar. 2010.

[92] B. Liu and H. Tian. A Bankruptcy Game-Based Resource Allocation Approach among Virtual Mobile Operators. *IEEE Communications Letters*, 17(7):1420–1423, July 2013.

[93] G. Zhang, K. Yang, J. Wei, K. Xu, and P. Liu. Virtual Resource Allocation for Wireless Virtualization Networks using Market Equilibrium Theory. In *IEEE INFOCOM Workshops*, pages 366–371, Apr. 2015.

[94] T. D. Tran and L. B. Le. Stackelberg Game Approach for Wireless Virtualization Design in Wireless Networks. In *IEEE International Conference on Communications (ICC)*, pages 1–6, 2017.

[95] E. Datsika, A. Antonopoulos, N. Zorba, and C. Verikoukis. Matching Game Based Virtualization in Shared LTE-A Networks. In *IEEE Global Communications Conference (GLOBECOM)*, pages 1–6, Dec. 2016.

[96] Y. Gu, W. Saad, M. Bennis, M. Debbah, and Z. Han. Matching Theory for Future Wireless Networks: Fundamentals and Applications. *IEEE Communications Magazine*, 53(5):52–59, May 2015.

[97] D. P. Bertsekas. Auction algorithms for network flow problems: A tutorial introduction. *Computational optimization and applications*, 1(1):7–66, 1992.

[98] S. Gold and A. Rangarajan. A graduated assignment algorithm for graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(4):377–388, 1996.

[99] A. E. Roth and M. Sotomayor. Two-sided matching. *Handbook of game theory with economic applications*, 1:485–541, 1992.

[100] J. W. Hatfield and P. R. Milgrom. Matching with Contracts. *American Economic Review*, 95(4):913–935, 2005.

[101] J. Pérez-Romero, J. Sánchez-González, R. Agustí, B. Lorenzo, and S. Glisic. Power-efficient resource allocation in a heterogeneous network with cellular and D2D capabilities. *IEEE Transactions on Vehicular Technology*, 65(11):9272–9286, 2016.

[102] P. Mary, M. Dohler, J.-M. Gorce, G. Villemaud, and M. Arndt. M-ary Symbol Error Outage over Nakagami-m Fading Channels in Shadowing Environments. *IEEE Transactions on Communications*, 57(10):2876–2879, 2009.

[103] G. Cocco, D. Gunduz, and C. Ibars. Throughput Analysis in Asymmetric Two-Way Relay Channel with Random Access. In *Proc. IEEE ICC*, pages 1–6, June 2011.

[104] G. Bianchi. Performance Analysis of the IEEE 802.11 Distributed Coordination Function. *IEEE Journal on Selected Areas in Communications*, 18(3):535–547, March 2000.

[105] D. Malone, K. Duffy, and D. Leith. Modeling the 802.11 Distributed Coordination Function in Nonsaturated Heterogeneous Conditions. *IEEE/ACM Transactions on Networking*, 15(1):159–172, 2007.

[106] 3rd Generation Partnership Project. LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Layer Procedures (3GPP TR 36.213 version 12.4.0 Release 12). Dec. 2014.

[107] A.R. Jensen, M. Lauridsen, P. Mogensen, T.B. Sørensen, and P. Jensen. LTE UE Power Consumption Model: For System Level Energy and Performance Optimization. In *Proc. IEEE VTC Fall*, pages 1–5, Sept. 2012.

[108] J.-P. Ebert, S. Aier, G. Kofahl, A. Becker, B. Burns, and A. Wolisz. Measurement and Simulation of the Energy Consumption of a WLAN Interface. *Telecommunication Networks Group, Technical University of Berlin*, June 2002.

[109] H. Schwarz, D. Marpe, and T. Wiegand. Overview of the Scalable Video Coding Extension of the H.264/AVC Standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(9):1103–1120, Sept. 2007.

[110] ITU-T and ISO/IEC (JVT). Joint Scalable Video Model 11 (JSVM version 9.19). March 2011.

[111] V. Rodoplu and T.H. Meng. Bits-per-Joule Capacity of Energy-Limited Wireless Networks. *IEEE Transactions on Wireless Communications*, 6(3):857–865, March 2007.

[112] N. Ding, D. Wagner, X. Chen, A. Pathak, Y. C. Hu, and A. Rice. Characterizing and Modeling the Impact of Wireless Signal Strength on Smartphone Battery Drain. *SIGMETRICS Performance Evaluation Review*, 41(1):29–40, June 2013.

[113] 3rd Generation Partnership Project. Technical Specification Group Radio Access Network, Spatial channel model for Multiple Input Multiple Output (MIMO) simulations (3GPP TR 25.996 version 12.0.0 Release 12). Sept. 2014.

[114] E. Datsika, A. Antonopoulos, N. Zorba, and C. Verikoukis. Adaptive Cooperative Network Coding Based MAC Protocol for Device-to-Device Communication. In *IEEE International Conference on Communications (ICC)*, pages 6996–7001, June 2015.

[115] A. Gumaste, T. Das, K. Khandwala, and I. Monga. Network Hardware Virtualization for Application Provisioning in Core Networks. *IEEE Communications Magazine*, 55(2):152–159, 2017.

[116] 3rd Generation Partnership Project. 3GPP Technical Report 36.300, Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN): Overall description, (Rel. 8). Oct. 2007.

[117] L. Chen, Y. Zhou, and D. M. Chiu. Smart streaming for online video services. *IEEE Transactions on Multimedia*, 17(4):485–497, 2015.

[118] Y. Ding, Y. Yang, and L. Xiao. Multisource Video On-Demand Streaming in Wireless Mesh Networks. *IEEE/ACM Transactions on Networking*, 20(6):1800–1813, Dec. 2012.

[119] M. Sikora, J. N. Laneman, M. Haenggi, D. J. Costello, and T. E. Fuja. Bandwidth- and power-efficient routing in linear wireless networks. *IEEE Transactions on Information Theory*, 52(6):2624–2633, 2006.

[120] F. Richter, A. J. Fehske, and G. P. Fettweis. Energy Efficiency Aspects of Base Station Deployment Strategies for Cellular Networks. In *IEEE Vehicular Technology Conference (Fall)*, pages 1–5, Sept. 2009.

[121] 3rd Generation Partnership Project. LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Radio frequency (RF) system scenarios (3GPP TR 36.942 version 13.0.0 Release 13). Jan. 2016.

[122] A. S. Kelso and V. P. Crawford. Job Matching, Coalition Formation, and Gross Substitutes. *Econometrica: Journal of the Econometric Society*, pages 1483–1504, 1982.

[123] G. Bolch, S. Greiner, H. de Meer, and K. S. Trivedi. *Queueing Networks and Markov Chains: Modeling and Performance Evaluation with Computer Science Applications*. John Wiley & Sons, 2006.

[124] J.R. Artalejo and A. Gómez-Corral. *Retrial Queueing Systems: A Computational Approach*. Springer, Berlin, 2008.

[125] YouTube. Live encoder settings, bitrates and resolutions. `https://support.google.com/youtube/answer/2853702`, 2017. Accessed on: 2018-07-01.

[126] Skype. How much bandwidth does Skype need? `https://support.skype.com/en/faq/FA1417/how-much-bandwidth-does-skype-need`, 2017. Accessed on: 2018-07-01.

[127] M. Jaber, Z. Dawy, N. Akl, and E. Yaacoub. Tutorial on LTE/LTE-A Cellular Network Dimensioning using Iterative Statistical Analysis. *IEEE Communications Surveys & Tutorials*, 18(2):1355–1383, Secondquarter 2016.

[128] R. Jain. *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling*. John Wiley & Sons, 1990.

[129] Z. Pi, J. Choi, and R. Heath. Millimeter-wave Gigabit Broadband Evolution toward 5G: Fixed Access and Backhaul. *IEEE Communications Magazine*, 54(4):138–144, Apr. 2016.

[130] A. Kaul, K. Obraczka, M. Santos, C. Rothenberg, and T. Turletti. Dynamically distributed network control for message dissemination in ITS. In *IEEE/ACM 21st Int. Symp. on Distributed Simulation and Real Time Applications*, 2017.

[131] NGMN Alliance. Guidelines for LTE Backhaul traffic estimation. `https://www.cisco.com/c/dam/en/us/solutions/service-provider/docs/backhaul-traffic.pdf`, July 2011. Accessed on: 2018-07-01.

[132] Qualcomm. Qualcomm Research LTE in Unlicensed Spectrum: Harmonious coexistence with WiFi. https://www.qualcomm.com/media/documents/files/lteunlicensed-coexistence-whitepaper.pdf, June 2014. Accessed on: 2018-07-01.

[133] 3GPP. Study on Licensed-Assisted Access to Unlicensed Spectrum; (Release 13), TR 36.889 V0.1.1,. Nov. 2014.

[134] N. Bhushan, J. Li, D. Malladi, R. Gilmore, D. Brenner, A. Damnjanovic, R. T. Sukhavasi, C. Patel, and S. Geirhofer. Network densification: the dominant theme for wireless evolution into 5G. *IEEE Communications Magazine*, 52(2):82–89, Feb. 2014.

[135] J. Puzzanghera. Net neutrality's repeal means fast lanes could be coming to the internet. Is that a good thing? http://www.latimes.com/business/la\-fi\-net\-neutrality\-fast\-lanes\-20171213\-story.html, Dec. 2017. Accessed on: 2018-07-01.