Optimizing routing and radio resource allocation for Multihop D2D Communications in 5G Networks

Safwan Alwan

UNIVERSITÉ PARIS-EST

Paris-Est Créteil Val-de-Marne University (UPEC)
Laboratory of Images, Signals and Intelligent Systems (LISSI)
MSTIC Doctoral School

Doctoral Thesis

By

Mr. Safwan ALWAN

Submitted to obtain the degree of

Doctor of Philosophy

From the University of Paris-Est

Optimizing Routing and Radio Resource Allocation for Multihop D2D Communications in 5G Networks

Publicly defended on the 4th of December 2019 before the following jury members

Dr. Lila Boukhatem	Reviewer	Associate Professor - HDR, Paris-Sud University, LRI Laboratory, Orsay, France.
Dr. Megumi Kaneko	Reviewer	Associate Professor - HDR, National Institute of Informatics (NII), Tokyo, Japan.
Dr. Cédric Adjih	Examiner	Researcher, National Institute for Research in Computer Science and Control (INRIA), Saclay, France.
Dr. Fabrice Guillemin	Examiner	Research Engineer, Orange Labs, Lannion, France.
Dr. Marcelo Dias de Amorim	Examiner	Research Director, French National Center for Scientific Research (CNRS), Paris, France.
Dr. Nathalie Mitton	Examiner	Research Director, National Institute for Research in Computer Science and Control (INRIA), Lille-Nord Europe, Lille, France
Dr. Ilhem Fajjari	Advisor	Research Engineer, Orange Labs, Châtillon, France
Prof. Nadjib Aitsaadi	Supervisor	Full Professor, UPEC/UVSQ Paris-Saclay, Guyancourt, France.

To my parents, my wife, my sisters and my brothers.

Acknowledgments

Foremost, I would like to express my sincere gratitude to my supervisor Prof. Nadjib Aitsaadi who has guided and taught me a lot throughout the journey of this Ph.D. thesis, which allowed me to learn and acquire the spirit of scientific research. Because the same thing can also be said about my advisor Dr. Ilhem Fajjari, I would like to thank her too for continuous support.

Besides, I would like to thank all jury members for offering their valuable time to review my thesis and examining my defense. Huge thanks are also due to them for valuable comments, questions, and insightful remarks.

My sincere thanks also go to the staff of the LISSI laboratory represented by the director Prof. Yacine Amirat and also to my teaching colleagues in the department of R&T in the IUT of Créteil/Vitry especially to Prof. Christian Lafont, the head of the department.

I would also like to thank my lab-mates for the discussions, the inspirational talks, and all the fun we have had together in the last four years. Among them, special thanks go to Roua Touihri, Fetia Bannour, Mohamed Alouane, Nicolas Khoury, and Arnaud Flori.

Last but not least, I would like to express my gratitude and love to my parents, Wadia and Galal, my wife, Ghita, and my sisters and my brothers for their love and support.

Abstract

Recently, Device-to-Device (D2D) has been brought inside mobile (cellular) networks with the introduction of the LTE-D2D standard into the 5G ecosystem. This cellular D2D operates in the same operator's frequencies used for regular communications with access points (i.e., base stations). In D2D mode, terminals can communicate directly and do not need to go through a base station. However, D2D communications are authorized and controlled by operators to implement their requirements and policies. A notable example of D2D is data offloading, which helps in reducing traffic congestion in mobile networks. In this scenario, terminals collaborate using their D2D connections to carry data, usually over multiple D2D hops, using other terminals as relays and avoiding base stations. However, the latter still must decide on routing (e.g., which devices should be part of the path) and wireless resource allocation (which frequencies to use by devices). Also, base stations must manage interferences between D2D and cellular communication since they all share the same spectrum. Besides, there is also the energy issue in employing battery-constrained terminals as relays. Another concern, in offloading designs, is how they scale when terminals density increases, such as in crowded-platform scenarios. These scenarios include mobile users in waiting halls of airports and train stations, or stadiums. In such situations, the decision problems mentioned before must be solved rapidly. Doing so avoids long delays in communications that can affect user experience or limit responsiveness. In this thesis, we address the problem of optimizing routing and wireless resource allocation in multihop D2D systems with a focus on data offloading. Our proposals to solve

the problem consider practical aspects of the LTE-D2D standard. Moreover, we also address the mentioned energy and scalability concerns. We propose three contributions to deal with these problems. In the first contribution, we propose a novel method (JRW-D2D) to solve jointly routing and resource allocation in the aim of offloading unicast flows inside one cell over the LTE-D2D relaying system. The proposal JRW-D2D is based on Integer Linear Programming (ILP) and gives good results in terms of reliability, latency, and acceptance ratio. In the second contribution, we present two methods to solve the same problem for both unicast and multicast traffic. In the first step, we introduce an optimal ILP-based method (JRW-D2D-MC) to solve routing and resource allocation jointly. Next, to address the scalability issue in JRW-D2D-MC, we propose another scalable method (JRW-D2D-CG) based on the Column-Generation technique. Finally, our third contribution considers the energy issue, where we put forward two energy-aware schemes to solve routing and resource allocation. Initially, we propose an ILP-based method for Energy-Efficient Joint Routing and Resource Allocation (JRRA-EE). In the next step, we highlight the non-scalability of JRRA-EE and introduce a novel parametric three-stage method called Heuristic Energy-aware Routing and Resource Allocation (HERRA). Both JRRA-EE and HERRA consider energy consumption using a state-of-the-art empirical model for LTE-D2D terminals. Moreover, we evaluate the performance of our contributions based on network simulations in NS-3, which we have extended to support the LTE-D2D standard.

Keywords:

5G, D2D communication, Routing, Radio Resource allocation, Data Offloading, Optimization.

Résumé

Récemment, D2D (Device-to-Device) a été intégré aux réseaux mobiles avec l'introduction de la norme LTE-D2D dans l'écosystème 5G. Ce D2D cellulaire fonctionne aux mêmes fréquences que l'opérateur utilisé pour les communications régulières avec les points d'accès (c.à.d. les stations de base). En mode D2D, les terminaux peuvent communiquer directement et n'ont pas besoin de passer par une station de base. Cependant, les communications D2D sont autorisées et contrôlées par les opérateurs pour mettre en œuvre leurs exigences et leurs politiques. Le délestage de données est un exemple notable de D2D, qui aide à réduire la congestion du trafic dans les réseaux mobiles. Dans ce scénario, les terminaux collaborent en utilisant leurs connexions D2D pour transporter les données, généralement sur plusieurs sauts D2D, en utilisant d'autres terminaux comme relais et en évitant les stations de base. Toutefois, ces derniers doivent décider du routage (par exemple, quels terminaux devraient faire partie du chemin) et de l'allocation de ressources sans fil (quelles fréquences à utiliser par les terminaux). De plus, les stations de base doivent gérer les interférences entre les communications, D2D et cellulaires, car elles partagent toutes le même spectre. En outre, il y a aussi le problème énergétique lié à l'utilisation de relais soumis aux contraintes de batterie. Un autre enjeu, dans ces conceptions de délestage, concerne la manière dont elles évoluent lorsque la densité des terminaux augmente (p. ex., dans les scénarios de plate-forme surpeuplée). Ces scénarios incluent des utilisateurs mobiles dans les halls d'attente des aéroports et des gares, ou des stades. Dans de telles situations, les problèmes de décision susmentionnés doivent être résolus rapide-

ment. Cela évite de longs délais dans les communications qui peuvent affecter l'expérience utilisateur ou limiter la réactivité. Dans cette thèse, nous abordons le problème de l'optimisation du routage et de l'allocation de ressources sans fil dans les systèmes D2D multi-sauts en mettant l'accent sur le délestage de données. Nos propositions pour résoudre le problème prennent en compte les aspects pratiques de la norme LTE-D2D. De plus, nous répondons également aux enjeux mentionnés en matière d'énergie et d'évolutivité. Nous proposons trois contributions pour traiter ces problèmes. Dans la première contribution, nous proposons une nouvelle méthode (JRW-D2D) pour résoudre conjointement le routage et l'allocation de ressources afin de délester des flux unicast sur un système de relais LTE-D2D. La proposition JRW-D2D est basée sur la programmation linéaire en nombres entiers (ILP) et donne de bons résultats en termes de fiabilité, de latence et de taux d'acceptation. Dans la deuxième contribution, nous présentons deux méthodes pour résoudre le même problème pour les trafics unicast et multicast. Dans la première étape, nous présentons une méthode optimale basée sur ILP (JRW-D2D-MC) pour résoudre conjointement le routage et l'allocation de ressources. Ensuite, pour résoudre le problème de non-évolutivité de JRW-D2D-MC, nous proposons une autre méthode évolutive (JRW-D2D-CG) basée sur la génération de colonnes. Enfin, notre troisième contribution aborde la question de l'énergie, dans laquelle nous avons présenté deux systèmes axés sur l'énergie pour résoudre les problèmes de routage et d'allocation de ressources. Dans un premier temps, nous proposons une méthode (JRRA-EE) basée sur ILP. Dans l'étape suivante, nous mettons en évidence la non-évolutivité de JRRA-EE et présentons une nouvelle méthode paramétrique à trois étapes appelée (HERRA) et basée sur l'heuristique. Les deux méthodes JRRA-EE et HERRA considèrent la consommation d'énergie à l'aide d'un modèle empirique de pointe pour les terminaux LTE-D2D. De plus, nous évaluons la performance de nos contributions sur la base de simulations de réseau dans NS-3 que nous avons étendu pour prendre en charge la norme LTE-D2D.

Mots-clès :

5G, Communication D2D, Routage, Allocation de ressources radio, Délestage de données, Optimisation.

Contents

Chapter 1

Introduction

The fifth-generation (5G) of broadband cellular technologies is a significant leap forward from 4G, which was revolutionary for its time. The current 4G networks cannot meet the future requirements of maintaining massively connected devices while providing low latencies and being very efficient in the frequency spectrum. Moreover, present-day data speeds and latencies offered by 4G are considered subpar for future applications and expected growth trends of wireless devices and mobile traffic. However, these requirements pose several challenges to overcome in 5G networks.

5G has to deal with the growing volume of data traffic, which has always been one of the major drivers behind the development of future network technologies. Factors behind this traffic increase include i) the proliferation of devices connected to the Internet ii) the increase in applications demands for high data volumes, and iii) the advent of new kinds of applications and services. The massive increase in the number of devices and connections, caused by the emergence of Internet-of-Thing (IoT), presents severe challenges to 5G, although they require only low-rate data transfers. To put things in perspective, the connected devices will reach 28.5 billion devices by 2022, up from 18

billion in 2017 with more than half of those devices and connections are of the Machine-to-Machine M2M type [1].

Furthermore, in pursuing solutions to the previous challenges, 5G has to i) increase the capacity while ensuring no or limited increase in CAPEX (CAPital EXpenditure), and support real-time data and high reliability for critical and emergency services, ii) support a wide range of 4G and post 4G air interface enhancements [2] (e.g., massive MIMO [3] and millimeter Wave (mmWave) antennas [4]), iii) cope with latency and bandwidth requirements of eXtended Reality (XR) services [5, 6], and iv) provide fast and scalable deployments for architectures with many layers of different connectivity. In light of network densification, this includes flexible architectural support for Heterogeneous wireless Networks (HetNet) [3] and Device-to-Device Communications (D2D) [7]. The normative 3GPP 5G specifications define requirements on capabilities performance targets that include, among others [8]:

i) Scalable support for network customization: using network slicing [9, 10] and Network Function Virtualization (NFV) [11],

ii) Ubiquitous connectivity support for fixed, mobile, wireless and satellite access technologies,

iii) Efficient use of resources, both in user and control planes, to support variable services which extend from low data-rate IoT [12] and vehicular communications [13, 14] to high-speed multimedia,

iv) Efficient utilization of the allocated spectrum,

v) Efficient energy and battery consumption for both infrastructure and end-user devices,

vi) Support for seamless mobility in densely-populated areas and heterogeneous environments, and

vii) Cellular coverage extension [15] via the cooperative relaying capabilities of enhanced user-devices.

From this standardizing work [8], one can see that i) high-speed data, ii) end–to-end latency less than 10 ms, and iii) ubiquitous connectivity targets are the notable characteristics of 5G networks that are expected to support a broad spectrum of applications and services.

1.1 Device-to-Device Communications (D2D)

Device-to-Device (D2D) communication is one of the key features of 5G networks. This concept refers to the ability of the user terminals to communicate directly in a peer-to-peer manner without the need to pass through an access point. Historically, the idea of exploiting peer-to-peer communications within cellular networks is not new. A theoretical architecture enabling multihop relays over mobile stations to the base station was given in academia in [16]. However, it is only recently that D2D has been considered to be integrated into the next generation networks after a plethora of scientific work identifying potential gains and use cases. As depicted in Figure 1.1, a distinct feature of D2D, when employed in cellular networks, is that infrastructure is involved in the assistance and coordination of the D2D control functions (e.g., resource allocation, routing, synchronization, session establishment, and authentication). Figure 1.1 also shows the three possible operation scenarios regarding the cellular coverage. Namely, i) In-Coverage Scenario, ii) Partial-Coverage Scenario, and iii) Out-of-Coverage Scenario.

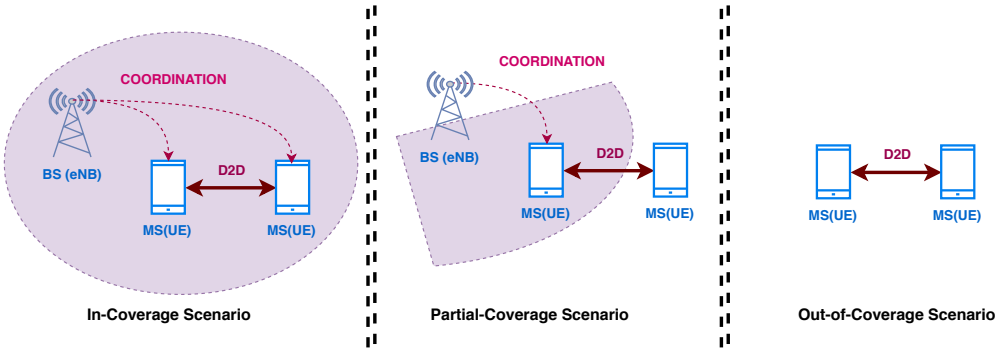


Figure 1.1 – D2D inside a cellular network: coverage scenarios

3GPP introduced D2D within LTE-A architecture in 3GPP Release 12 as an enabler for Proximity Services (ProSe) which included two essential functions: i) direct discovery, and ii) direct communication between (enhanced) User-Equipments (UEs). The primary motivation behind this LTE-D2D standard, aka LTE-Direct, is to provide competitive wireless technology for public safety networks to be used by first responders. In addition to public safety applications, 3GPP ProSe supports discovery-based services for commercial use cases and network coverage extension using UE-to-Network relay. Figure 1.2 shows the evolution of the D2D support in the 3GPP standards. In 3GPP Release 13, LTE-

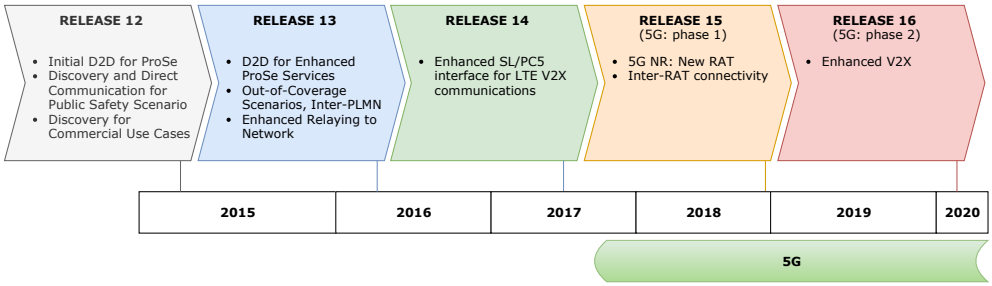


Figure 1.2 – State of the D2D support in LTE releases

D2D improved the ProSe support for various scenarios including inter-operator and out-of-coverage scenarios. A significant enhancement in the standard was introduced in Release 14. It consists of supporting Vehicle-to-Everything (V2X)

communications. In Release 15, which represents the first phase of 5G, the support for IoT and Wearables was included in the LTE-D2D. In fact, it was decided that IoT and wearable devices would benefit from short D2D links and optimized UE-to-Network relays through cooperating UEs, to extend their battery life. In Release 16, the support for V2X was further enhanced.

As a final note on terminology, the term D2D includes, in addition to LTE-D2D, other peer-to-peer wireless standards such as ZigBee, Near Field Communications (NFC), Bluetooth, and Wi-Fi Direct.

1.2 Advantages and Use Cases of D2D

Several potential benefits of using D2D communications within cellular networks have been identified when implemented properly [17]. These advantages include:

i) *Efficient use of the operator's spectrum*: The additional layer of D2D communications can be configured to reuse the same spectrum of the cellular communication layer yielding high efficiency of frequency reuse of the whole cellular/D2D system.

ii) *Energy efficiency*: The short-range aspect of D2D communications may result in less energy consumption from the device's viewpoint.

iii) *Low latency*: In addition to being short-range, direct communication between devices without an intermediate access point will offer very low communication delays.

iv) *Enabling new services and use cases*: Due to D2D being inherently local and proximity-aware, incorporating D2D in cellular networks allows new services and use cases, as explained in the following paragraphs.

1.2.1 Public-Safety Use Cases

For a better assessment of emergencies, first-responders (e.g., police, fire, and medical emergency services), need broadband access in the next-generation public safety network [18]. Using D2D technology, public safety networks enable terminals to communicate directly without any support from the infrastructure while being scalable to substantial group calls [19]. It is worth noting that the standard LTE-D2D was initially developed for public safety use cases to provide the necessary functionalities: Push-To-Talk (PTT), Direct communications between terminals and Group communications [20].

1.2.2 Locality and Context-Aware Services

Reliable discovery of nearby devices, using the D2D protocol, enables various use cases and services. Both fixed and mobile devices, (e.g., infrastructure sensors, beacons of transport and businesses, mobiles, and tablets) can interact with each other to provide locality and context-aware services. Typical examples include: i) *social discovery applications:* e.g., finding nearby friends of persons with mutual interests in Facebook or LinkedIn, ii) *local guidance and advertisement:* e.g., searching for nearby bus stations, ATMs, restaurants, and museum guidance, and iii) *transport information:* e.g., notification of the arrival of the next bus, parking availability.

1.2.3 Local Content Sharing

UEs can use their D2D interfaces to exchange files rapidly while consuming lower energy than the conventional method involving the cellular connection. These interfaces also facilitate streaming video locally between users by forming clusters. Moreover, social applications can make use of D2D capabilities to

share content between users in proximity.

1.2.4 Network Range Extension

A UE can reach a cellular BS through one or more UEs serving as relays to the network. An example of this scenario is devices that are either in weak connectivity areas (e.g., indoor or cell edge) or devoid of enough power to reach a distant BS (e.g., smartwatch). A neighboring UE with satisfactory connectivity or sufficient power source can connect with those devices in its vicinity, using its D2D interface, and forwards, then, their data to the BS through its cellular interface.

1.2.5 Traffic (Data) Offloading

D2D can be employed to enhance the networking of the future 5G networks in several ways. One way is to offload the traffic [21] from the cellular infrastructure to the direct communication between two nearby UEs, which discover each other using D2D-based discovery protocol [22]. Data offloading techniques [23] efficiently deal with the problem of congestion in next-generation cellular networks. In this context, a congestion-prone BS may take advantage of a secondary wireless technology to offload the circulating traffic between UEs and thus saving resources and bandwidth. A D2D-based protocol can provide the secondary mechanism to carry the offloaded traffic [24]. Besides, the D2D communications can be either in the operator's band (i.e., using 3GPP LTE-D2D) or in the unlicensed spectrum (e.g., based on Wi-Fi Direct [25]). The offloaded traffic can be either unicast or multicast, following BS-to-UE(s), UE-to-BS, or UE-to-UE models. Moreover, the offloading can be achieved using a multihop network of D2D links.

1.2.6 IoT and V2X Communications

Due to ultra-low latency requirements, D2D-based solutions offer scalable and resilient support for Machine-to-Machine (M2M) and Machine-Type Communications (MCT), including IoT and Vehicular communications (V2X) scenarios [26, 27]. The latter includes Vehicle-to-Vehicle (V2V), Vehicle-to-Infrastructure (V2I), and Vehicle-to-Pedestrian (V2P).

1.3 D2D Architecture

1.3.1 Spectrum Allocation

Figure 1.3 shows the architecture of D2D communications within the cellular infrastructure. The direct communication between devices is made possible by introducing a new kind of lateral link between devices in addition to the conventional wireless links: DownLink (DL) and UpLink (UL). From the spectrum allocation viewpoint, wireless D2D technology can operate in the same band as cellular technology or another band. In the former configuration, D2D is said to be In-Band (IB). The latter configuration is called Out-Of-Band (OOB), in which D2D uses another band (usually unlicensed one). In the in-band configuration, the shared band can be either i) orthogonally allocated to both communication types, called In-Band Overlay (IBO), or ii) totally reused by both types at the same time, called In-Band Underlay (IBU) mode. The orthogonality in IBO configuration can also be achieved using Time-Division (IBU-TD); that is, the spectrum is used in alternating periods between the two communication types.

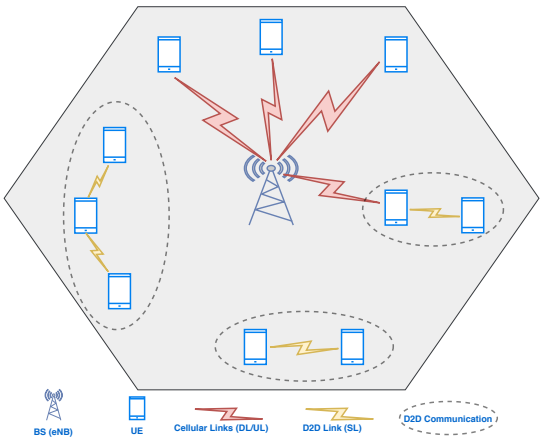


Figure 1.3 – D2D within cellular communications

1.3.2 3GPP Architecture for D2D (LTE-D2D)

To support LTE-D2D, an enhanced user equipment (UE) implements an additional protocol stack besides the conventional one. This new LTE-D2D stack provides the so-called Proximity-based Services (ProSe) to the upper layer(s) [28]. ProSe includes: i) *Direct Discovery:* a service whereby a UE can detect and identify other UEs in its proximity, ii) *Direct Communications:* UEs can directly communicate with each other bypassing the cellular infrastructure, and iii) *UE–to-Network Relay:* remote UE uses another UE as a relay in the network.

From an upper-level perspective, ProSe is carried over a new type of wireless link beside the conventional ones: i.e., DownLink (DL), and UpLink (UL). This lateral link between UEs is called SideLink (SL). In LTE-D2D, SL is configured to use the **same frequency resources as UL** to increase the overall spectral efficiency [29]. It also **reuses much of UL structure and hardware** to add another efficiency dimension. From the lower layers perspective, SL presents its direct communication services to the upper layers in terms of no-feedback SL Radio Bearers (SLRBs). This is done to present uniform support for both unicast and multicast IP communications [28].

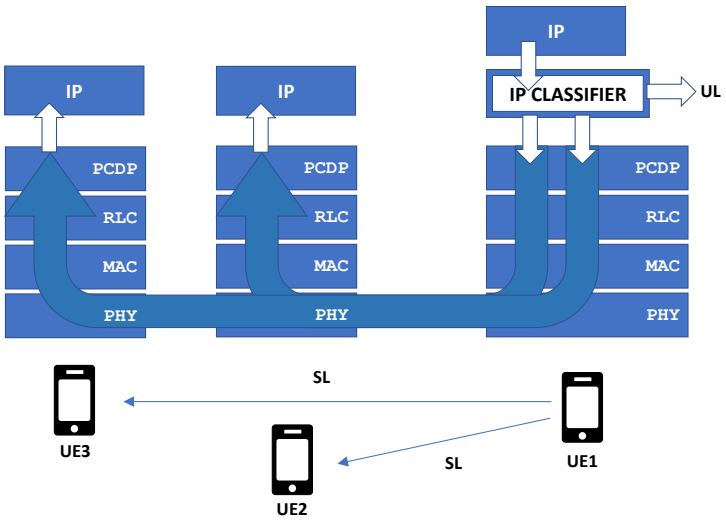


Figure 1.4 – LTE-D2D protocol stack for Direct Communications

1.3.3 LTE-D2D Protocol Stack for Direct Communications

Figure 1.4 depicts the LTE-D2D stack to support direct communications. Hereafter, we provide a brief top-down description.

PDCP/RLC

Similar to their counterparts in the conventional LTE communication stack, Packet Data Convergence Protocol (PDCP), and Radio Link Control (RLC) layers provide IP packet segmentation, header compression, and security procedures. A single SLRB is identified by a pair of PDCP/RLC entities connected in tandem at the source UE and the corresponding pair(s) at the destination UE(s). At the interface with incoming packets from the upper IP layer resides an IP-flow classifier that directs each IP packet to its corresponding SLRB PDCP/RLC entities.

MAC

The Medium Access Control (MAC) layer serves the upper layers by transmitting Transport Block (TB) composed of RLC Protocol Data Units (PDU) from

possibly several SLRB bearers as long as they have the same destination. Each TB is identified by its layer L2 identifiers, namely, i) source *ProSe-UE-ID*, and ii) the destination *ProSe-L2-Destination-ID*. A TB is transmitted when a new SL transmission opportunity arrives while Hybrid Automatic ReQuest (HARQ) operations in LTE-D2D are restricted to blind retransmissions (i.e., with no feedback) to increase the reliability. Hence, each TB is further retransmitted **three times** in the subsequent transmission opportunities with different redundancy versions.

PHY

Similar to UL, SL transmission, at the Physical (PHY) layer, uses the Single Carrier Orthogonal Frequency Modulation (SC-OFDM) format using the grid of resource blocks (RBs). The latter occupies a subframe, i.e., a Transmissions Time Interval (TTI), which lasts 1 ms and is characterized by a bandwidth of 12 subcarriers (180 kHz) in the frequency domain. However, unlike UL, SL allocations are organized in longer periodic intervals called SideLink Control Periods (SC-Periods), which can be configured between 40 and 320 subframes in length. As depicted in Figure 1.5, a SC-Period starts with a control part followed by a data part. However, the information, on which subframes and RBs are available for the operation, is conveyed by a configuration parameter called a *resource pool*.

A UE interested in SL reception scans continuously the configured resource pool(s) (the control part of SL periods) to check for incoming data. On the other hand, a UE wanting to transmit on SL may be configured to go ahead and autonomously selects a RB subset among a resource pool configured for this mode of operation. These resources are used to transmit the UE's data. Even with in-coverage scenarios, the base-station, also known as eNodeB (or eNB) in LTE, may configure this autonomous mode for its UEs. Another possible op-

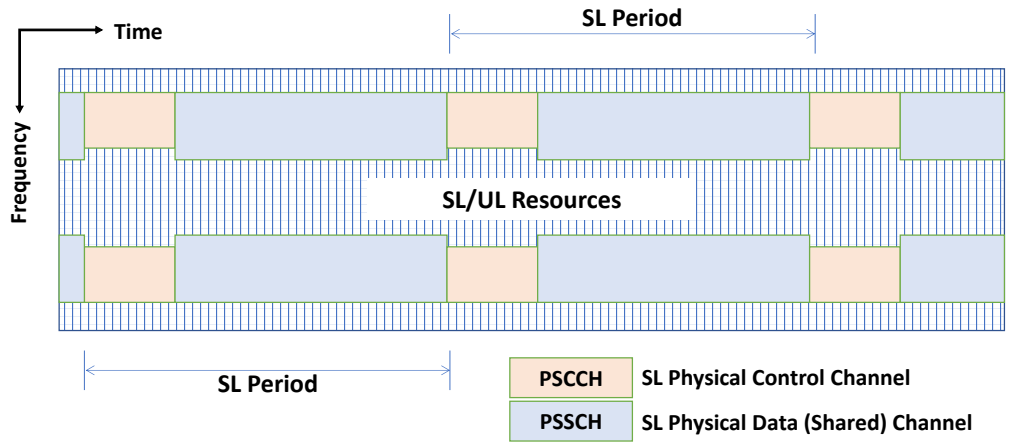


Figure 1.5 – SL channel structure in LTE-D2D (direct communications only)

tion is to configure *scheduled-resources pool* to be used in the scheduled mode, which gives the eNB a finer control over resource allocation. In this mode, a grant from the eNB to the UE determines which RBs and subframes to be used by the UE to transmit on SL. Throughout this thesis, we assume the latter mode of resource allocation. In doing so, the eNB has total control over the resource allocation.

Synchronous Operation of SL

Transmissions and receptions on SL are synchronous. This means that they must have a common synchronization reference for all parties in the system. With in-coverage scenarios, where a UE is inside the coverage zone of an eNB, the UE synchronizes its SL operation to the timing of the related macro-cell, which acts as a synchronization reference. Further procedures and provisions are given in the standard, allowing some UEs to relay timing reference to extend the synchronization zone even under out-of-coverage scenarios [28].

Half-Duplex Operation of SL

As per LTE-D2D standard, the duplex mode of SL is half-duplex, meaning that a UE can not simultaneously both listen and transmit on SL. However, the rules for role-switching are not specified and left to the application under consideration. Also, UEs connected to the eNB (i.e., macro-cell) are required by the standard to give their UL transmissions higher priority over SL transmissions since both compete for the same Single Carrier Orthogonal Frequency Division Multiple Access (SC-OFDMA) transmitter. So, whenever there are UL data or reports, e.g., Sounding Reference Signal (SRS), Channel Quality Indicator (CQI), etc.; an ongoing SL subframe, if any, must be dropped. These properties make SL transmissions more opportunistic and intermittent and less reliable than UL.

1.4 D2D Challenges

Despite its numerous advantages, D2D raises several concerns and technology challenges:

1.4.1 Peer Discovery

To enable discovery-based services, a D2D-enabled UE has to discover other UEs in its proximity [30]. In principle, there are two modes of discoverability: i) closed (or restricted) mode and ii) open mode. In restricted discovery, a UE can only be discovered with the explicit permission of end-users. In open discovery, a UE implicitly agrees to be found by any other UE in its vicinity. The main problems are i) how to increase the probability and speed of the discoverability of UEs [31], and, ii) how to optimize the required resources and power [32]. Given the mobility aspect and the possible involvement of the cellular system,

the responses to this challenge incorporate i) the design of discovery signals (i.e., beacons) and their periodicity, ii) the amount and location of the physical resources allocated to the discovery operation, and iii) usage of measurements by the base-stations to optimize the process.

1.4.2 Mode Selection

The direct D2D link between the UEs may not always be the best choice to exchange data. Mode selection [33] refers to the problem of selecting the proper mode (D2D or cellular) for communication to achieve a given performance objective. The last can be a user-centric objective [34] or system-centric objective [35]. Mode selection can be made by the cellular network (i.e., centralized) or the UEs themselves (i.e., distributed). Possible performance targets include i) enhancing the spectral efficiency, ii) minimizing the power, and iii) improving the information rate.

1.4.3 Resource Allocation and Interference Management

An active D2D link requires frequency resources (i.e., Physical RBs) to perform the transmission. In light of the co-existing cellular system, these resources may be shared with ordinary cellular communications (e.g., UL communications). Since the spectrum is a scarce resource, allocating sufficient resources to both types of communications is extremely challenging [36], especially for in-band (IBO/IBU) D2D configurations. In the orthogonal schemes (IBO), a common design challenge consists in finding the optimal division of spectrum (i.e., RBs) between the two tiers of communication to reach some performance target. For reuse schemes (IBU), a similar challenge arises as to which RBs are shared or reused to keep interference below a certain threshold. In both cases,

the interference management is a crucial part of the design. The purpose of managing the interference is to avoid harmful interferences between the two link types (D2D and cellular) and among D2D links themselves. Out-of-band systems (D2D) also may benefit from the coordination of the cellular system to minimize the interference's effect between D2D with the other unlicensed-band technologies.

1.4.4 Routing over D2D Links

The second tier of D2D communications is by no means limited to single-hop connections. A network of UEs may be employed to deliver data over multiple hops. This scheme applies to the different traffic models inside the cellular systems i) unicast, multicast, or broadcast, and ii) BS-to-UE or UE-to-UE. This multihop mechanism is particularly interesting for offloading scenarios where the traffic may be delivered over a network of UE acting as cooperative relays. The problem of finding the proper routes over D2D networks can be solved using three methods. In addition to the classical ad-hoc methods, the routing can i) be handled centrally by the BS (i.e. eNB), or ii) be performed by the UEs with coordination from the BS. Routing over D2D topologies should consider their particularities [37], specifically: i) nodes mobility, ii) co-existing cellular communications, and iii) intermittent aspect of the D2D interface. Moreover, the overall system can achieve additional gains when the routing procedure is paired with resource allocation and interference management. In such an approach, the routing decisions (network layer operations) can interact with resource allocations (MAC layer operations) and interference measurements (Physical layer procedures). Such joint treatment is considered as a cross-layer optimization technique [38]. The latter represents a departure from the strict OSI model.

In addition to the previous challenges, D2D has other concerns that include: i) security and authentication, ii) coordination between operators, iii) privacy and lawful interception, iv) charging and business models, and motivating users to allow their devices to act as relays for others.

1.5 Problem Statement

In this thesis, we propose to push the D2D technology in LTE-A networks beyond the capabilities of single D2D hops. To this end, we aim to employ a multihop system of D2D-enabled UEs (e.g., mobile phones, PC-attached modems, access relays), to deliver traffic between the UEs themselves. Hence, the traffic is offloaded without the involvement of the eNB except for the management and the coordination of radio resources. Besides, we note that such networks might have dynamic topologies because nodes are mobile, and links can appear or disappear.

The research problem of this thesis is to propose new cellular traffic offloading schemes relying on routing methods and radio resource allocation algorithms. These methods are supposed to support the offloading operation mentioned above. Moreover, the resource allocation is based on LTE-A Orthogonal Frequency Division Multiple Access (OFDMA) explained earlier. Such proposed schemes aim to achieve the following objectives:

i) Optimizing the use of (shared) frequency spectrum which is a rare resource,

ii) Ensuring communication quality in terms of throughput, delay, packet loss rate.

iii) Minimizing interference with cellular communications,

iv) Optimizing routing between source and destination(s), and

v) Conforming to the LTE-D2D standard.

Furthermore, the proposed schemes should take into account i) computation time and memory usage, ii) signaling volume (i.e., the number of subframes exchanged to make decisions), iii) various types of traffic (e.g., unicast or multicast, constant or variable bit-rate), iv) energy consumption of battery-limited UEs, and v) scalability in dense UEs scenarios (i.e., in waiting halls of airports and train stations, or stadiums).

1.6 Contributions

In this thesis, we put forward our three main contributions to solve the routing and wireless resource allocations in multihop LTE-D2D communications within LTE-A cellular systems. These contributions are summarized in the following paragraphs.

- Our first contribution[1] addresses the joint routing and OFDMA resource allocation problem in D2D networks to offload unicast UE-to-UE traffic. To this end, we advance a formulation for the problem as Mixed Integer Linear Programming (MILP). To solve the MILP model, we propose a Branch-and-Cut method, named Joint Routing and Wireless Resource Allocation for multihop D2D communications (`JRW-D2D`). The model takes into account factors that limit spectrum reuse as well as other LTE-D2D technology constraints such as half-duplex operation and contiguity in resource block allocations. To evaluate our proposal, we have implemented the LTE-D2D protocol stack in the NS-3 network simulator, which

[1] The results of this contribution were published in [39].

lacks in support of LTE-D2D protocol. Network simulations show that our method JRW-D2D yields excellent results in terms of reliability, latency, and the ratio of offloaded flows compared to other basic one-sided optimal strategies, i.e., that optimize only routing or resource allocation, including an interference-aware heuristic scheme.

- In the second contribution[2]; we extend the scope of the first contribution to include a uniform traffic model that supports both unicast and multicast traffic types. Moreover, we also address the scalability issue in solving routing and resource allocation jointly in LTE-D2D multihop networks. To do so, we formulate the joint problem as an Integer Linear Programming (ILP) model, which considers, as before, the factors that limit spectrum reuse as well as the LTE-D2D limitations: namely, the half-duplex mode and contiguity of RBs. Then, we put forward a novel two-stage algorithm, named Joint Multicast Routing and Wireless allocation in D2D communications (JRW-D2D-MC). The devised algorithm consists of an initial stage that prefilters the flows that can be routed considering the current state of the network, to reduce the size of the ILP model. Subsequently, JRW-D2D-MC makes use of the celebrated Branch-and-Cut algorithm to solve the reduced ILP model. Network simulations in NS-3 augmented with our-home-grown D2D module shows that JRW-D2D-MC is excellent in terms of flow acceptance rate and latency.

  However, to address the scalability concern in the initial formulation, JRW-D2D-MC, we propose a novel path-based ILP formulation in which a routing tree is formulated in terms of its constituent paths. Moreover, for reason of speed, we propose a sub-optimal solution method, named JRW-D2D-CG, based on the Column-Generation framework with a pricing

[2]The preliminary results of the second contribution were published in [40]

problem. The latter allows us to consider only paths that are likely to enhance the solution. We adjust the pricing problem to be more tractable, and then, we use a fast algorithm based on the Bellman-Ford algorithm to find advantageous paths. Based on extensive network simulation in the NS-3 environment, we show that our novel proposal `JRW-D2D-CG` achieves good performances in terms of reliability, latency, and scalability.

- The third contribution[3] of this thesis addresses the energy issue in the relaying D2D network. To this end, we present two approaches for the eNB to optimize the centralized decision problem of routing and RB allocation. In the first approach, we present an ILP-based formulation for the problem that considers the realistic LTE-D2D capabilities and constraints as before. To assess our proposal's effectiveness, we run out network-level simulations based on our NS-3 module, as described earlier. Extensive simulations show that `JRRA-EE` is better compared to other one-sided optimal strategies, including an energy non-aware variant, in terms of i) network lifetime, ii) packet loss, and iii) service interruption rate.

  However, despite these advantages, a downside of `JRRA-EE` is that it does not scale well with high-density topologies. For this reason, we present another scalable approach, named Heuristic Energy-aware Routing and Resource Allocation (`HERRA`), which consists of a parametric three-stage method. Performance evaluation, using network simulations in NS-3, shows that our new proposal, `HERRA`, outperforms the initial `JRRA-EE` in the matter of convergence time. Owing to massive speedups, up to six orders of magnitude, `HERRA` scales very well in denser topologies while having some performance gaps, especially in terms of packet loss.

[3] Preliminary results were published in [41]

1.7 Thesis Outline

The following material in this manuscript is organized as follows. In Chapter 2, we present a survey on related work concerning routing and resource allocation in D2D communication. In Chapter 3, we introduce the first thesis contribution on *Joint Unicast Routing and Wireless Resource Allocation in Multihop LTE-D2D Communications*. Chapter 4 presents the second contribution on *A Scalable Joint Routing and Resource Allocation Scheme: D2D-based Unicast and Multicast Data Offloading*. In Chapter 5, we present the third contribution on *D2D-Based Cellular Traffic Offloading: An Energy-Aware Scalable Heuristic Scheme*. Finally, Chapter 6 presents a conclusion of this thesis, providing insights into our current and future work.

Chapter 2

Related Work

D2D wireless communications in cellular networks is an extremely challenging paradigm that has aroused the interest of both industry and academia. In this section, we summarize the most relevant related work that helped us to have an insight into the multihop D2D routing and resource allocation problems. We also summarize the most relevant approaches found in the literature about routing and content-delivery in the context of multihop in D2D systems. Many works in the literature address the peer-to-peer communication between User Equipments (UEs), which is, on its own, a quite old idea. Nevertheless, we focus here on the D2D communication in LTE-A cellular systems, which is relatively a new concept. In these systems, the control plane resides in the eNB while the data plane is offloaded to (a network of) UEs. However, one should note that D2D communication is used sometimes as a generic term that also includes other peer-to-peer access technologies relying on Wi-Fi and Bluetooth.

2.1 Literature on Unicast D2D Systems

In [42], the authors propose an algorithm for dynamic UE relay selection to assist in delivering the BS-to-UE traffic. The algorithm is a distance-based heuristic that aims to keep the signaling and feedback overhead at an acceptable level by limiting the number of candidates relays for the targeted UE. Using numerical simulations, the authors show that the presented algorithm significantly reduces the overhead without compromising system performance. The work considers multiple-BSs scenarios and both IBO and IBU modes of D2D. However, by design, the authors only consider a two-hop system that incorporates one D2D link.

In [43], the authors consider the optimal transmission scheduling and congestion control in multihop D2D communications that underlie cellular networks. They consider: i) interference condition for the D2D and the conventional cellular modes, and ii) the QoS requirements of each traffic flow. Their formulation employs the Lyapunov optimization theory and considers the following problems: i) end-to-end rate control, ii) joint routing and channel assignment, and iii) power allocation, to solve the global problem using a sub-optimal approach. The proposed approach remarkably considers also the stability of queues in the forwarding UEs because of the dynamic nature of the routing employed. In other words, the algorithm solves for optimal routing on a per-time-slot basis. Notably, the algorithm also makes some assumptions relevant to LTE-D2D: i) D2D links share the uplink's spectrum, and ii) half-duplex nature of the D2D transmissions. Nevertheless, the presented method does not allocate the spectrum in resource-block granularity, which makes it less practical in the context of LTE-D2D.

In [44], the authors propose a scheme to employ D2D multihop communications in cellular networks for public safety scenarios under partial cellular

coverage, which is a typical use case during disasters. Based on their implementation of a system-level simulator of the 3GPP LTE-D2D standard, the authors have demonstrated improvements in energy and spectral efficiency when compared to conventional communications. However, there is no routing algorithm presented in this work since it only employs predefined routes from a far-away UE to reach an operating base-station passing by other relaying UEs.

In [45], the authors put forward a two-stage method to find multihop D2D paths under a limit on the maximum interference incurring at conventional mobile users. Based on numerical simulations, significant improvements in throughput can be achieved using multihop paths compared to single-hop D2D communication. However, the proposed method is highly generic. Indeed, only one single assumption is considered by the authors to apply their approach: downlink resources are shared by D2D and conventional communication.

In [46], the authors propose a D2D-assisted relaying system to offload the BS-UE traffic to secondary BSs in HetNets. The overall system is composed of a Main BS (MBS) and several secondary Small BSs (SBS). Within this system, a UE may connect directly to the MBS or via another relay UE connected to a secondary SBS. In other words, the two-hop relay sub-system incorporates one BS-UE link and another UE-UE D2D link. Using a dynamic tri-partite graph-based formulation, the authors formulate the problem of maximizing the number of connected UEs to the system. To this end, the authors put forward an algorithm that decides the optimal UE-UE and relay-SBS associations, (i.e., a form of routing), based on a 0-1 ILP model. The presented algorithm uses dynamic programming to solve the ILP model as the problem is proven to be NP-hard. Through numerical simulation, the authors show that their proposal improves the offloading capacity outperforming related schemes. They also report en-

hancements in average UE energy consumption. Thanks to its genericity, the presented design is applicable in LTE-D2D systems, although it does not consider the LTE-D2D specificities. However, this may not yield the optimal performance, e.g., regarding the use of spectrum, since the design abstracts channels as monolithic where the two-tier of communications use orthogonal bands.

The authors in [47] propose an offloading scheme based on multi-RAT D2D communications, including the unlicensed RATs (e.g., Wi-Fi, Bluetooth), to achieve high link spectrum efficiency. The presented scenario is a single-hop UE-UE, where the BS takes charge of selecting the best D2D interface between the UEs. Using the framework of stochastic geometry, the authors formulate the problem of maximizing link spectrum efficiency using the retention probability as a parameter. The authors assume that the licensed D2D RAT (e.g., LTE-D2D) operates using the IBU mode. However, the resource allocation is abstracted as monolithic channel access. The routing is limited to the selection of a single-hop interface for each UE-UE pair. Based on numerical simulations, the authors demonstrate significant improvements in link spectrum efficiency and the coverage probability compared with the traditional non-offloaded scheme.

In [48], the authors develop a scheme of Quality-of-Service (QoS) provisioning, in terms of statistical delay-bound, for the D2D-based BS-to-UE traffic offloading. The presented system employs single UE-to-UE hop in addition to the BS-to-UE relaying connection. The authors aim to maximize the effective global capacity of the two-hop system with statistical bound on the delay QoS. The authors demonstrate, through numerical simulation, that their scheme is capable of achieving the indicated goal. In light of LTE-D2D, the formulation is a bit generic but applicable. However, the authors do not address the resource allocation, and the spectrum allocation is abstracted in terms of a (monolithic) dedicated frequency channel. Moreover, the coexistence between conventional

and D2D communications is guaranteed by assuming the OOB mode.

In [49], the authors present a theoretical framework to evaluate the energy and spectral efficiency in large-scale mobile cellular traffic offloading systems based on D2D operating in the IBO and IBU modes. Since the results are derived using closed-form analytical expressions, the authors believe that those results present practical tools for the design and the evaluation of future D2D-enabled cellular networks. Based on these analytical results, the author outlined an optimal spectrum partitioning scheme networks operating in IBO mode. The objective is to maximize the network's energy and spectral efficiency with constraints on the user outage and the D2D transmitters' power. The numerical results, confirming the analytical ones, suggest that the IBU mode is more spectrum and energy-efficient than the IBO. However, the addressed traffic models are UE-to-UE, unicast, and broadcast traffics along with the cellular one. The authors address no specific LTE-D2D constraints, and moreover, they assume that D2D shares the spectrum with the DL, which contradicts the current LTE-D2D's reality.

2.2 Literature on Multicast D2D Systems

In [50], the authors give an insightful study of resource allocation for multicast wireless OFDMA-based systems. This work covers various aspects of channel-aware resource allocation of wireless multicast systems as well as multicast-related concepts such as group formation, single-rate, and multi-rate transmissions. However, the authors address only multicast downlink transmissions. Stated differently, the authors consider those systems where data transmissions start at the base station, and the UEs, in the multicast groups, may act as forwarders if needed.

In [51], the authors propose a D2D-based offloading strategy for BS-to-UEs traffic (multicast or broadcast). The scheme employs an initial stage where the BS uses conventional multicast communication to reach a group of (seed) UEs, which have favorable channel conditions. Next, the seed UEs use their opportunistic D2D interfaces to diffuse the content to the rest of the UEs to complete the dissemination. Another stage of unicast transmission may be needed to reach those UEs not served by the previous steps. The authors propose a central algorithm, to be run in the BS, based on the Reinforcement Learning (RL) framework to control the operation. This algorithm decides which UEs should act as seeders (i.e., they are served through the cellular multicast) and which should be served using opportunistic D2D. A generic OOB D2D technology is assumed to cooperate with the standard LTE mobile system supporting the multicast service. Network simulations based on NS-3 demonstrate that D2D allows optimizing the multicast communication saving up to 90% of the BS's radio resources.

In [52], the authors study the multi-copy data dissemination in mobile opportunistic Delay-Tolerant Networks (DTN). In such networks, the content delivery may take up to days. The authors propose a probabilistic delay-constrained formulation to determine the optimal multicast graph that minimizes the communication cost. Then, they propose two algorithms: centralized and distributed. They evaluate the performance under the random walk mobility model and real-world mobility traces.

In the same vein as [52], the authors in [53] propose a multicast architecture for the D2D content delivery in cellular networks. In the proposed architecture, the content originates at the base station, and a one-hop multicast relaying is employed to deliver the content. However, the authors focus primarily on the mode selection (i.e., cellular, or D2D) for content delivery and the caching strat-

egy.

In [54], the authors examine the problem of power minimization in multicast multihop D2D networks through user grouping strategies. In this work, the authors propose two greedy sub-optimal algorithms to work around the NP-completeness of the problem. Nonetheless, the authors limit the scope to a single content delivery that begins at the base station. Moreover, the proposed schemes make very general assumptions about the underlying D2D technology used to offload content delivery. Besides, the schemes do not deal with the problem of resource allocation.

2.3 Literature on Other Routing Models in D2D Systems

In [55], the authors present a generic routing and resource allocation scheme based on multihop D2D for M2M communication. The system aims to improve the end-to-end connectivity between (MCTD-) UEs inside an LTE-A cell where the traffic starts in sensing nodes and ends in a collector node. The authors use an acyclic directed graph to model the routing process for this UEs-to-collector-UE traffic where the collector UE is the root of the graph. The proposed route selection is a distributed task, and the intermediate nodes in the graph aggregate the data received from their predecessors, including their data. These intermediate relays, then, direct the traffic towards the collector node. A simple RB allocation approach is outlined by the author to allocate RBs for nodes in proportion to their relative closeness to the collector node. The authors assume an IBO mode of operation and recognize the half-duplex aspect of D2D. However, no performance validation or evaluation is given in this article.

2.4 Literature on Energy-Aware D2D Routing

The authors of [56] demonstrate that LTE-D2D cooperative relays can save significant amounts of energy when compared to conventional Base Station (BS) to UE communications. Besides, the authors put forward a collaborative relaying design intending to increase the UE's battery life. Their approach seeks to maximize the utilization of UEs possessing high energies to carry the traffic of those with low power. Their numerical simulations reveal that their method decreases the outage probability of the cellular cooperating UEs.

In [57], the authors put forward another scheme to deliver BS-to-UE video content by a cooperative D2D multihop routing. The proposed system employs a generic framework to avoid disruption caused by the depletion of D2D UE's energy budget. Their algorithm seeks to optimize the budget utility by the joint scheduling of the routes and traffic workloads according to the energy efficiency of every D2D link.

In [58], the authors propose an energy-efficient routing protocol in WiFi-Direct cluster-based networks. The designed protocol borrows ideas from the well-known protocols LEACH and HEED from the wireless sensor networks (WSN). Using numerical simulations, the authors demonstrate that their scheme considerably saves the network's energy when compared to the usual peer-to-peer mode of WiFi-Direct.

In [59], the authors introduce a heuristic algorithm for the energy-efficient routing for UE-UE unicast traffic. Both channel reuse and power allocation are jointly undertaken to achieve satisfactory performance. The simulations demonstrate significant improvements in the energy-efficiency of the multihop D2D communication systems.

2.5 Comparative Summary and Remarks on Literature

In Table 2.1, we summarize the reviewed literature on routing and resource allocation in D2D systems. The table highlights the presented proposals according to different criteria. In particular, in regards to traffic/route models, proposals are classified according to:

i) the end-to-end traffic model: i.e., unicast, multicast, or broadcast.

ii) the path (route) model: i.e., BS-UE (between the BS and UE in both directions), UE-to-UE, or BS-to-UE. Note that the BS-to-UE route model involves one (initial) cellular hop (i.e., DL), which is followed by D2D hops.

Moreover, we classify the algorithms of routing and resource allocation into centralized, executed by a single entity (i.e., BS), and distributed (i.e., UEs execute parallel tasks). As for interference management, we classify presented systems according to the spectrum coexistence models presented in Chapter 1, where:

i) IBO/IBU represents in-band schemes, where D2D uses the same (licensed) frequency band as the cellular communication in orthogonal (overlay) and non-orthogonal (underlay) forms.

ii) OOB represents out-of-band schemes, where D2D uses another (usually unlicensed) band. These systems virtually employ different non-cellular radio technology (e.g., Wi-Fi Direct).

Besides, we classify the proposals according to how they abstract the physical channel access, which ranges from: i) Logically-Abstracted, ii) Single (Monolithic) Channel, iii) Multiple Orthogonal (Non-Overlapping) Channels, up to iv)

Resource Block (RB)-Level. The RB-level modeling of the resource allocation provides the greatest flexibility because links can be allocated a different number of RBs according to different traffic and QoS requirements. The multiple-channels abstraction, where the total band is divided into fixed spectrum width channels, is less flexible abstraction than the RB-level one but more flexible than the single-channel abstraction.

We remark that the research work that addresses the joint routing and resource allocation, in LTE-based infrastructures, is limited despite the abundant related proposals about D2D in general.

Table 2.1 – Comparative summary of literature on routing and resource allocation in D2D within cellular systems

Paper	Context	Proposal	Formulation	Targets	Traffic/Route Models	Routing	Resource Allocation	Interference Handling	Physical Resource Abstraction	Channel Models	Evaluation Metrics	Evaluation Method
[42]	Relaying/Offloading	Routing	Heuristic	Min. Signaling Overhead	Unicast BS-UE	Centralized	N/A	IBO-TD/IBU	Single Channel	Large-Scale Loss Exponent (+ Rayleigh fast-fading)	Average Sum-Rate Outage Probability	Numerical Simulation
[43]	Relaying/Offloading	Routing, Channel Assignment Power Allocation	Lyapunov Optimization	Optimize Scheduling Congestion Control	Unicast UE-to-UE	Centralized	Centralized	IBU	Channels	Large-Scale Loss Exponent (+ Rayleigh Fast Fading)	Rejection Rate, Throughput Backlog Size	Numerical Simulation
[44]	Coverage Extension	Evaluation Framework	N/A	Proof-of-Concept	Unicast BS-UE	Predefined	Predefined	IBU	RBs	WINNER	Energy Efficiency Spectral Efficiency	Numerical Simulation (System-Level)
[45]	Relaying	Routing, Power Allocation	Heuristic (Dijkstra-Based)	Max. Throughput	Unicast UE-to-UE	Centralized	N/A	IBU	Single channel	Large-Scale Loss Exponent	Throughput	Numerical Simulation
[46]	Offloading	Routing Mode Selection	Graph-Based ILP	Max. № of Connected UEs	Unicast UE-to-UE	Centralized	Centralized	IBO/OOB	Single channel	Large-Scale Loss Exponent	Offloading Efficiency (UEs), Average Energy	Numerical Simulation
[47]	Offloading	Routing scheme RAT Selection	Stochastic Geometry	Max. Spectrum Efficiency	Unicast UE-to-UE	Centralized	Centralized	IBU/OOB	Single Channel	Large-Scale Loss Exponent (+ Rayleigh Fast Fading)	Spectral Efficiency, Coverage Probability	Numerical Simulation
[48]	Offloading	Power Allocation	analytical	Max. Capacity	Unicast, BS-to-UE (one D2D hop)	N/A	N/A	OOB	Single Channel	AWGN Channel, Nakagami-m Channel	Capacity	Numerical Simulation
[49]	Offloading	Evaluation Framework	analytical	Max. Energy-Spectral Efficiency	Unicast, Broadcast UE-to-UE	N/A	N/A	IBO/IBU	Single Channel	Large-Scale Loss Exponent (+ Rayleigh Fast Fading)	Energy-Spectral Efficiency	Numerical Simulation
[51]	Offloading	Routing	Reinforcement Learning	Min. № of RBs	Multicast BS-to-UEs	Centralized	N/A	OOB	RBs	Cost 231, Extended Pedestrian A	№ of RBs Used Packet Delivery Ratio	Network Simulation (NS-3)
[52]	Content Dissemination	Routing	Heuristic (Graph-Based)	Min. Communication Cost	Multicast UE-to-UE	Centralized, Distributed	N/A	OOB	Single Channel	N/A	Cost, Delay Success, Delivery Rates	Numerical Simulation
[53]	Relaying/Offloading	Routing (Mode selection)	Heuristic	Content-Delivery	Multicast BS-to-UEs	Centralized	N/A	OOB	Single Channel	N/A	Serving Time Delivery Ratio	Numerical Simulation
[54]	Relaying/Offloading	Routing (Cluster Formation)	Heuristic	Min. Transmission Power	Multicast BS-to-UEs	Centralized	N/A	OOB	Single Channel	Large-Scale Loss Exponent (+ Log-Normal Shadowing)	Power Consumption, Delivery Ratio	Numerical Simulation
[55]	M2M	Routing Resource Allocation	Graph-Based	Enhance End-to-End Connectivity	Unicast (Many-to-One) UEs-to-UE	Distributed	Abstracted	N/A	Abstracted	N/A	Average Sum-Rate Outage Probability	N/A
[56]	Relaying/Offloading	Routing (Cluster Formation)	Heuristic	Max. Battery Lifetime	Unicast,UE-to-BS (one D2D hop)	Centralized	Abstracted	IBO/IBU	Abstracted	WINNER II (Indoor for D2D)	Relaying Time	Numerical Simulation (Event-Driven)
[57]	Relaying/Offloading	Routing (+Cashing Strategy)	Heuristic	Min. D2D Outage	Unicast UE-to-UE, BS-to-UE	Distributed	N/A	IBO-TD	Single Channel	Large-Scale Loss Exponent	Throughput	Numerical Simulation
[58]	Relaying/Offloading	Routing (Cluster Formation)	Heuristic	Max. Network Lifetime	Unicast UE-to-BS	Distributed	N/A	OOB	Single Channel	N/A	Energy Dissipation	Network Simulation (NS-2)
[59]	Relaying/Offloading	Routing	Heuristic	Max. Energy-Efficiency	Unicast UE-to-UE	Centralized	N/A	IBU	Single Channel	Large-Scale Loss Exponent	Energy-Efficiency, Average Hop-Count	Numerical Simulation

Chapter 3

Joint Unicast Routing and Wireless Resource Allocation in Multihop LTE-D2D Communications

Contents

5G aims to maximize the data rate and to handle the billions of video, voice, data, and IoT flows. For this reason, the macro-cells will be very congested and may fail to satisfy the end-users. In this context, the data offloading scheme is conceived to route intra-cell traffic among the D2D-enabled user equipments reusing wireless uplink resources and thus increasing the overall spectral efficiency. In this chapter, we address the joint routing and OFDMA resource allo-

cation problem in the D2D network. To do so, first, we formulate the problem as Mixed Integer Linear Programming. The model takes into account factors that limit spectrum reuse as well as other LTE-D2D technology constraints such as half-duplex operation and contiguity in resource block allocations. Then, we propose a novel scheme named Joint Routing and Wireless allocation in D2D communications (JRW-D2D), which is based on the branch-and-cut algorithm. In order to gauge the effectiveness of our proposal, we implement the standard LTE-D2D protocol stack, including our scheme JRW-D2D, in the NS-3 network simulator. The results obtained are very promising in terms of reliability, ratio of admitted D2D flows and latency in comparison to other basic one-sided optimal strategies including an interference-aware heuristic scheme.

3.1 Introduction

THERE is no denying that the Fourth Generation (4G) of mobile cellular network, Long Term Evolution (LTE), held the promise of higher data rate and enhanced the Quality of Service (QoS). But, the growth of video-centric and social media services has led to the explosion of traffic demand. In addition, the Internet of things will exponentially increase the number of flows in the cellular network. Consequently, the current cellular infrastructures struggle to accommodate the required network resources and link capacities. This trend is set to continue, and recent statistics highlight that the number of connected devices is estimated to reach 50 billion by 2020 while the mobile data traffic is expected to grow to reach 49 exabytes per month by 2021 [60].

Therefore, discussions of a new standard have taken place in both industry and academia to design the Fifth Generation (5G) mobile cellular network architecture. The main objective of 5G is to ensure the QoS satisfaction of

the different applications and to deal with diverse deployments in terms of available resources and connected devices requirements. In this context, 5G puts forward disruptive technologies making use of i) massive MIMO [61] and millimeter-Wave antenna systems [62], ii) Multiple Radio Access Technologies (Multi-RAT) [63, 64], iii) small cells deployment [65], and iv) advanced Device-to-Device (D2D) communications. All these techniques aim to increase the capacity of networks in order to handle a large number of connections and data volume at high throughput and very low latency.

The main idea behind D2D is to enable direct communications between devices in close proximity and thus to bypass macro base-stations. D2D was incorporated in LTE-A to increase the spectral efficiency of cellular systems and to support new use cases such as i) public safety scenarios, ii) device-discovery for commercial applications, iii) D2D-network relays, etc. D2D is also one pillar of 5G architecture, enabling operators to ensure extended and controlled connectivity while reducing the network's cost thanks to the traffic offloading solutions. In doing so, the data plane is moved from the operator's infrastructure (i.e., E-UTRAN, and EPC) to end-users' devices (i.e., UE). However, the control plane is managed by the operator and hosted in E-UTRAN. This will alleviate the infrastructure's load while enabling large numbers of simultaneous connections with better QoS.

D2D raises several design challenges [17, 66], such as coexistence with conventional communications mode (macro-cell), spectrum reuse and resource allocation, mode-switching, extending single-hop scenarios to multihop ones, etc. In this chapter, we address the routing and wireless resource allocation problems in D2D communications. Multihop D2D seeks to enhance the utility of D2D systems by increasing the communication range and reducing the load in the operator's infrastructure. Multihop D2D system must adopt various poli-

cies with respect to the routing, resource allocation, interferences: intra-mode (i.e., D2D links) and inter-mode (i.e., D2D, and conventional communications). Note that a sidelink communication (i.e., D2D) uses the same physical resource (transceiver and spectrum) of uplink communication. That means that UE cannot simultaneously do both sidelink and uplink communications. In addition, UE cannot simultaneously transmit and receive in the sidelink. Consequently, each link in the D2D path is half-duplex, and only non-critical (in terms of latency and bandwidth) traffic can be handled. We formulate the joint routing and resource allocation problem of D2D communications while considering: i) contiguity of OFDMA resource block allocation, ii) interference, and iii) half-duplex mode of operation in LTE-D2D as a Mixed Integer Linear Programming (MILP) problem. The objective is to maximize the bandwidth of each flow (i.e., best-effort).

Concerning the related work, in this chapter, we address the joint optimization of resource block allocation and routing for multihop communications. Unlike [43], which is the closest one to our proposal in this chapter, we adopt a semi-static routing where path establishment takes into account the current state of interfering links, but the path is held for the whole period of communication. We also model the allocation problem to the resource block level taking into account the fact that they are allocated in a contiguous manner (3GPP uplink constraint). Besides, we notice that existing literature on multihop D2D communications shows varying degrees of relevancy to LTE-D2D standard and lack of proposal validation using network simulators due to the support for D2D standards. To cope with this limitation, we implemented in NS-3 the full 3GPP LTE-D2D protocol stack to evaluate the performance of our proposal.

To solve the above problem, we propose a novel scheme, based on the branch-and-cut algorithm [67], named **Joint Routing and Wireless allocation in D2D**

communications (JRW-D2D). In this chapter, we assume a dense deployment of UEs in a delimited area, such as a stadium. Consequently, the UEs are not mobile. It is worth noting that the routes set up for flows are *semi-static* paths. In other words, each path is maintained for the whole period of communication to avoid excessive signaling to reconfigure D2D links. On the other hand, resource allocations are dynamically executed every assignment interval to cater for flow's arrivals and departures.

To assess the performance of our proposal JRW-D2D, we implemented the LTE-D2D protocol stack in the NS-3 network simulator to support this standard. In doing so, the whole protocol stack is simulated, and hence the conclusions will be more significant than the numerical simulations. The results obtained demonstrate the effectiveness of JRW-D2D in terms of the optimality, the ratio of admitted D2D flows, and latency. In addition, we compared our proposal to other basic one-sided optimal strategies and an interference-aware heuristic scheme. A one-sided optimal strategy is one that is optimal only in one sense, either in terms of routing or in terms of resource block allocation.

The remainder of this chapter is organized as follows. In section 3.2, we will describe in detail our system model for the offloading application and how we formulate the decision problem as a MILP model that includes routing and resource block allocation. Then, in section 3.3, we will describe our proposal JRW-D2D used to solve the underlying problem. Next, in section 3.4, we will present our evaluation methodology and network simulation results. Finally, Section 3.5 will conclude the chapter.

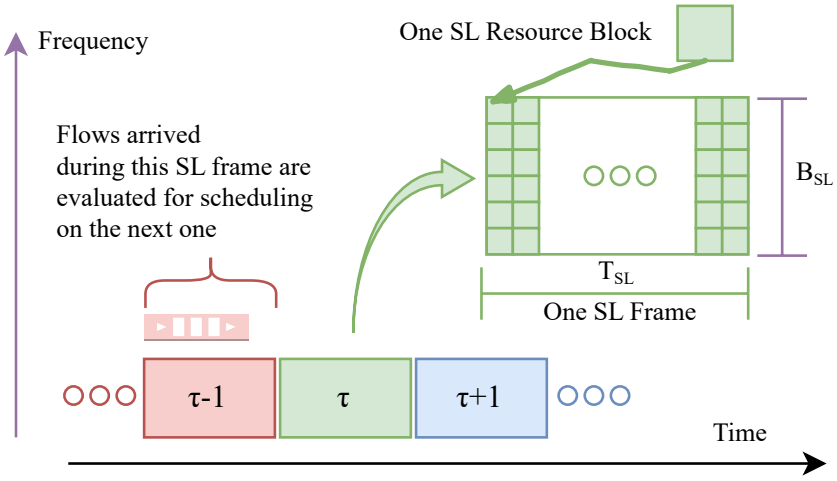


Figure 3.1 – Sidelink frame structure and scheduling

3.2 System Model and Problem Formulation

Our system model considers N UEs inside the coverage zone of a single LTE-A eNB. These UEs, which supposed to support the LTE-D2D protocol, are willing to offload the intra-cellular traffic between them when commanded to do so by the central controller in the eNB. We also assume that these UEs are quasi-stationary nodes. The eNB supervises the offloading operation over this D2D network by continuously allocating radio resources in every SL frame with decision instants given by:

$$t = \tau \times T_{SL} \qquad \text{for } \tau = 0, 1, 2, 3, \ldots$$

where T_{SL} is the duration of SL frame. The SL frame, or the SL control period in LTE-D2D terminology, is the scheduling time unit in SL, which spans multiple one-millisecond time slots (i.e., multiple TTIs). Figure 3.1 illustrates the structure of SL frame. The eNB models the D2D topology as a symmetric directed graph $\mathbb{G} = (\mathcal{V}, \mathcal{E})$. The set of vertices $\mathcal{V}$ and the set of edges $\mathcal{E}$ represent the UE

nodes and the links between the UEs (i.e. SLs), respectively. Note that a link in topology is formed, and hence an edge exists in $\mathbb{G}$, only when the achieved SNR is higher than a threshold γ_{TOPO}. This means that $\mathbb{G}$ is not connected, in the general case, and can be expressed as a union of connected sub-components: $\mathbb{G} = \mathbb{G}^1 \cup \mathbb{G}^2 \cup \ldots \cup \mathbb{G}^C$.

The problem of finding an offloading path for a flow $f^k \in \mathscr{F}$, whose source and destination are s^k, d^k respectively, can be formulated as follows: We introduce for each link e_{ij} a binary variable x_{ij} to indicate whether it is selected to be a part of some route. We also introduce for each node v_n a binary variable A_n^k that indicates whether it is associated with the flow f^k. In this formulation the offloading path for f^k is defined by the set of $\mathbb{P}^k \subseteq \mathscr{E}$:

$$\mathbb{P}^k = \left\{ e_{ij} \in \mathscr{E} \mid x_{ij} = 1 \wedge \mathrm{A}_i^k = \mathrm{A}_j^k = 1 \right\} \tag{3.1}$$

However, in order for equation (3.1), to meaningfully define a path, the solution space must respect some constraints defined in the following.

First, we impose that nodes are exclusive for concurrent flows. In other words, a node can route at most one flow at a time. Formally, this constraint is introduced as:

$$\forall v_n \in \mathcal{V}, \sum_{f^k \in \mathscr{F}} \mathrm{A}_n^k \leq 1 \tag{3.2}$$

In addition, if a node is associated with some flow it must have exactly one *incoming* link selected except at the source where there is none. This is formally imposed as:

$$\forall v_n \in \mathcal{V}, \sum_{e_{ij} \in \mathscr{E} \mid j=n} x_{ij} = \sum_{f^k \in \mathscr{F}, v_n \neq s^k} \mathrm{A}_n^k \tag{3.3}$$

Similarly, if a node is associated with some flow it must have exactly one *outgo-*

ing link selected except at the destination where there is none, or:

$$\forall v_n \in \mathcal{V}, \sum_{e_{ij} \in \mathcal{E} | i=n} x_{ij} = \sum_{f^k \in \mathcal{F} | v_n \neq d^k} \mathrm{A}_n^k \tag{3.4}$$

Also, to ensure that node association is consistent with link selection, the following constraint imposes that the ends of a selected link are associated with the same flow:

$$\forall f^k \in \mathcal{F}, \forall e_{ij} \in \mathcal{E}, x_{ij} - 1 \leq \mathrm{A}_j^k - \mathrm{A}_i^k \leq 1 - x_{ij} \tag{3.5}$$

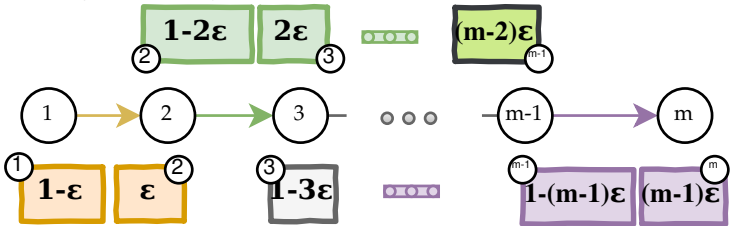
It is straightforward to see that, if some flow f^k is decided to be admitted, which is indicated by $\mathrm{A}_{s^k}^k = 1$, then $\mathbb{P}^k$, as defined in equation (3.1) and under Constraints (3.2) to (3.5), must contain only one simple path, which starts from s^k to d^k. However, these constraints do not rule out superfluous links (and nodes) from appearing in $\mathbb{P}^k$ forming *simple isolated* cycles between them.

To exclude these isolated simple cycles from the decision space, we propose the *token-split* method. If a pair of nodes (v_i, v_j) is selected to form a route, a token of one unit, $t_{ij}+t_{ji}=1$, is unconditionally split between them such that v_i and v_j receive t_{ij} and t_{ji} respectively. Note that the only constraint on t_{ij}, t_{ji} is that they are nonnegative reals. However, to exclude the above-stated cycles, we impose that each node must receive a total amount of tokens that is *strictly less* than 1. To see how this works, suppose that we have a cycle of m nodes, and m links selected. Then, the total tokens to be split among them equals exactly m. In this case, it is impossible to find a way to split tokens between consecutive pairs in the loop, such that each node receives *strictly less* than 1. To formulate such strict inequality by a non-strict one, a threshold parameter $0 < \epsilon < \frac{1}{2}$ may

be used. Then, the no-loop constraint can be stated as:

$$\forall v_n \in \mathcal{V}, \sum_{e_{ij} \in \mathcal{E} | i=n} t_{ij} \leq 1 - \epsilon$$

However, we must also be sure that such restriction does not rule out arbitrary paths in the solution space. Suppose that we have a path of $m \geq 3$ nodes with $m-1$ links selected. Then, we show that it is possible to split the total $m-1$ tokens respecting the previous constraints if $\epsilon \leq \frac{1}{m}$. To prove this, we can split the tokens such that the first $m-1$ nodes receive exactly $1-\epsilon$ token each, and as a consequence, the last one receives $(m-1)-(m-1)(1-\epsilon)$ or $(m-1)\epsilon$ token. This is explained graphically as follows:



To respect the no-cycle condition at the last node, we have $(m-1)\epsilon \leq 1-\epsilon$ which implies $\epsilon \leq \frac{1}{m}$ which completes the proof. To sum it all, if we set the parameter $\epsilon = \frac{1}{|\mathcal{V}|}$, where $|\mathcal{V}|$ is the total number of nodes, then all possible loops are excluded from the solution space without excluding any possible (simple) path from a source to a destination. Formally, the no-cycle constraints are given by:

$$\forall e_{ij} \in \mathcal{E}, x_{ij} + x_{ji} \leq 1 \tag{3.6}$$

$$\forall e_{ij} \in \mathcal{E}, t_{ij} + t_{ji} = x_{ij} + x_{ji} \tag{3.7}$$

$$\forall v_n \in \mathcal{V}, \sum_{e_{ij} \in \mathcal{E} | i=n} t_{ij} \leq 1 - \frac{1}{|\mathcal{V}|} \tag{3.8}$$

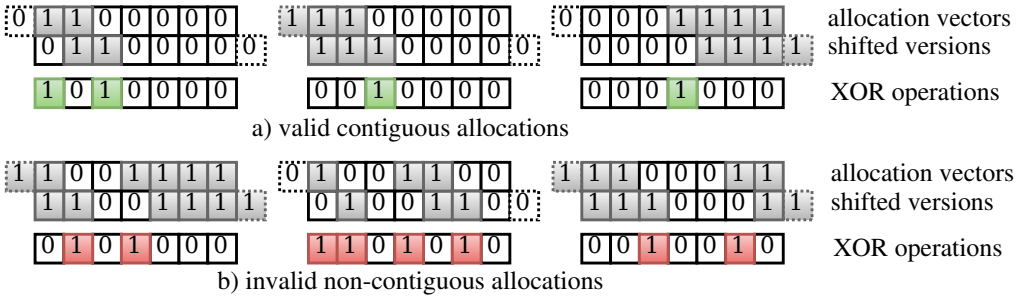
Given the half-duplex mode hardware constraint in D2D, UEs cannot simultaneously transmit and receive on SL. Therefore, active nodes must switch back

and forth between roles. In order to reduce the end-to-end delay, we require all non-successive nodes in a path to transmit in one period while their respective partners are listening to them, and in the next period, they swap roles. This principle of operation forces the links along a path to be scheduled in an alternating manner. The net effect of these assumptions is that the SL scheduler switches every SL frame between two sets of active UEs in order to maintain the ongoing flows. In other words, the active nodes $\mathcal{V}_{\mathrm{H}} \subseteq \mathcal{V}$ are divided into two sets: $\mathcal{V}_{\mathrm{H}}^{0} = \{v_n \in \mathcal{V} \mid \mathrm{H}_n = 0\}$ and $\mathcal{V}_{\mathrm{H}}^{1} = \{v_n \in \mathcal{V} \mid \mathrm{H}_n = 1\}$ where H_n are binary variables attached to the nodes. Hence a pair of nodes, having an active link between them, cannot be in the same *half-duplex set* (period):

$$\forall e_{ij} \in \mathscr{E}, x_{ij} \leq \mathrm{H}_i + \mathrm{H}_j \leq 2 - x_{ij} \tag{3.9}$$

In addition to its assigned half-duplex period, a transmitting node also needs frequency resources. In line with LTE-D2D standard, we assume a Frequency Division Duplex (FDD) cellular network where we assign a bandwidth $\mathrm{B_{SL}}$, composed of Ω contiguous OFDMA RBs, to the SL operation. Note that only contiguous RB allocations are feasible within this bandwidth because the SL has the same communication proprieties as the UL [68]. We represent the allocated RBs for a node v_n, by a vector of 0-1 variables R_n^{ω} for $\omega = 1, 2, \cdots \Omega$, where the variable R_n^{ω} indicates whether the RB number ω is allocated to v_n. To formulate the contiguity constraints, we use the *Hamming distance*. The Hamming distance $d_{\mathrm{H}}(\mathrm{V}_1, \mathrm{V}_2)$ between the vectors V_1 and V_2 is the number of positions at which the two vectors differ. The Hamming distance between two 0-1 vectors, $\mathrm{V}_1 = [\mathrm{V}_1^1, \mathrm{V}_1^2, \cdots, \mathrm{V}_1^n]^{\mathrm{T}}$ and $\mathrm{V}_2 = [\mathrm{V}_2^1, \mathrm{V}_2^2, \cdots, \mathrm{V}_2^n]^{\mathrm{T}}$, is the sum of component-wise XOR operation between the vectors (i.e. $d_{\mathrm{H}}(\mathrm{V}_1, \mathrm{V}_2) = \sum_{i=1}^{n} \mathrm{V}_1^i \oplus \mathrm{V}_2^i$). To check an allocation vector $[\mathrm{R}_n^1, \mathrm{R}_n^2, \cdots, \mathrm{R}_n^{\Omega}]^{\mathrm{T}}$ for contiguity, we remark that the Hamming distance between $[\mathrm{R}_n^1, \mathrm{R}_n^2, \cdots, \mathrm{R}_n^{\Omega-1}]^{\mathrm{T}}$ and its shifted version $[\mathrm{R}_n^2, \mathrm{R}_n^3, \cdots, \mathrm{R}_n^{\Omega}]^{\mathrm{T}}$ is

less than 2 if the $R_n^1 = 0$ and is less than 1 if $R_n^1 = 1$ as illustrated as follows:



Formally, this constraint is expressed as:

$$\forall v_n \in \mathcal{V}, \sum_{\omega=1}^{\Omega-1} R_n^{\omega} \oplus R_n^{\omega+1} \leq 2 - R_n^1 \tag{3.10}$$

In addition, RBs are allocated for some node v_n only when the node is a transmitter (i.e., one of its outgoing links is selected). Formally:

$$\forall v_n \in \mathcal{V}, \sum_{\omega=1}^{\Omega} R^{\omega} \leq \sum_{e_{ij} \in \mathcal{E} | i=n} x_{ij} \tag{3.11}$$

To reuse the spectrum efficiently, and to reduce power consumption, we require that nodes are not allocated RBs beyond the request of the associated flow or formally:

$$\forall f^k \in \mathcal{F}, \forall v_n \in \mathcal{V}, \sum_{\omega=1}^{\Omega} R_n^{\omega} \leq \Omega + \left(D^k - \Omega\right) A_n^k \tag{3.12}$$

Note that the relation between flow bit-rate R^k and the respective demand for RBs D^k is defined by [68] as:

$$R^k = \frac{\mathrm{TBS}\left(\mathrm{MCS}, D^k\right)}{T_{\mathrm{TTI}}} \quad [\mathrm{bps}] \tag{3.13}$$

where TBS is the MAC Transport Block Size function in bits as defined in [68] considering a baseline *Modulation and Coding Scheme* (MCS) for the SL and

T_{TTI} is the transmission time interval (i.e., the duration of a subframe) which is equal to 1 ms.

In the face of the reuse of RBs, system performance is limited by the interference caused by nodes transmitting using the same RB. To deal with interference, we assume a fixed power density scheme for the D2D emission. According to this scheme, the total emission power $S_{tx,n}$ of a node is proportional to the number of allocated RBs $\sum_{\omega=1}^{\Omega} R_n^{\omega}$. Formally,

$$S_{tx,n} = \Psi_{t,n} \cdot \sum_{\omega=1}^{\Omega} R_n^{\omega} \qquad [\text{mW}] \tag{3.14}$$

Furthermore, we assume a common emission power density, $\Psi_{t,n}$ [mW/RB], for all the D2D nodes (i.e., $\forall v_n \in \mathcal{V}, \Psi_{t,n} = \Psi_t$). Following the same *per-RB* treatment and assuming *flat block-fading* channel model, the overall Signal-to-Interference-plus-Noise Ratio (SINR) on the link e_{ij} is equal to:

$$\gamma_{ij} = \frac{g_{ij}\Psi_{t,i}}{\sum\limits_{v_n \in \mathcal{V}} g_{nj}\Psi_{t,n} + \Psi_{\sigma}} \tag{3.15}$$

where Ψ_{σ} and $\Psi_{t,n}$ represent the spectral densities (per RB) of the thermal noise and the transmission from v_n, and g_{ij} is the channel gain between the node pair (v_i, v_j).

Furthermore, additional variables $R_n^{\omega,0}$ and $R_n^{\omega,1}$ are defined to indicate whether the RB ω is used by v_n in $\mathcal{V}_H^0$ or $\mathcal{V}_H^1$ (i.e. half-duplex set of frames), respectively. Formally,

$$\forall (v_n, \omega) \in \mathcal{V} \times [1, \Omega], \quad R_n^{\omega,0} \triangleq R_n^{\omega} - R_n^{\omega,1} \tag{3.16}$$

$$\forall (v_n, \omega) \in \mathcal{V} \times [1, \Omega], \quad R_n^{\omega,1} \triangleq H_n \cdot R_n^{\omega} \tag{3.17}$$

An additional set of link-level auxiliary 0-1 variables are introduced as follows:

$$\forall (e_{ij}, \omega, p) \in \mathscr{E} \times [1, \Omega] \times \{0,1\}, \quad \mathrm{R}_{ij}^{\omega,p} \triangleq \mathrm{R}_{i}^{\omega,p} \tag{3.18}$$

$$\forall (e_{ij}, \omega, p) \in \mathscr{E} \times [1, \Omega] \times \{0,1\}, \quad \phi_{n,ij}^{\omega,p} \triangleq \mathrm{R}_{n}^{\omega,p} \cdot \mathrm{R}_{ij}^{\omega,p} \tag{3.19}$$

where $\mathrm{R}_{ij}^{\omega,p}$ indicates if the RB ω is used for the scheduled link e_{ij} during the p^{th} half duplex set, $\phi_{n,ij}^{\omega,p}$ is an interference indicator between node v_n and link e_{ij} on the RB ω.

To adhere to a linear formulation, further steps are needed to linearize the XOR-terms in Constraint (3.10) and the product terms in Constraints (3.17), (3.18) and (3.19).

We make use of a standard technique to linearize each XOR-term $x \oplus y$ by introducing an additional auxiliary 0-1 variable $\lambda_{xy}^{\oplus}$ and adding four more linear constraints as follows:

$$(\lambda_{xy}^{\oplus} \geq x-y), (\lambda_{xy}^{\oplus} \geq y-x), (\lambda_{xy}^{\oplus} \leq x+y), (\lambda_{xy}^{\oplus} \leq 2-x-y) \tag{3.20}$$

We use another standard technique to linearize each product term $x \cdot y$ by introducing an additional auxiliary 0-1 variable $\lambda_{xy}^{\odot}$ add four more linear constraints as follows:

$$(\lambda_{xy}^{\odot} \leq x), (\lambda_{xy}^{\odot} \leq y), (\lambda_{xy}^{\odot} \geq x+y-1) \tag{3.21}$$

To optimize the performances by minimizing interferences, SINR must be upper-bounded by a common threshold γ. To formulate this constraint on RB allocations, we translate this limit (i.e., SINR $\leq \gamma$) into the inequality $\mathscr{N} + \mathscr{I} \leq \mathrm{P}_r / \gamma$ where P_r is the received power.

$$\forall (e_{ij},\omega,p)\in\mathscr{E}\times[1,\Omega]\times\{0,1\},\quad \Psi_{\sigma}\mathrm{R}_{ij}^{\omega,p}+\sum_{n\neq i}g_{nj}\Psi_{t}\cdot\phi_{n,ij}^{\omega,p}\leq\frac{g_{ij}\Psi_{t}}{\gamma}\mathrm{R}_{ij}^{\omega,p} \tag{3.22}$$

As stated before, the function of our eNB is to schedule the SL resources in order to support the ongoing (already-admitted) flows and to handle newly-arriving flows trying to admit some of them when possible. In doing so, the objective is to maximize the overall utilization of system resources (nodes and RBs) while serving the maximum possible number of flows. To reach such objective, our utility function can be decomposed into three goals: i) maximizing the total number of allocated RBs, ii) maximizing the number of admitted flows, and iii) minimizing the total hop-count of the reserved paths.

We propose to formulate these goals as single objective-function of weighted-sums to complete the MILP formulation, developed so far, as follows:

$$\max_{\substack{x_{ij},\mathrm{A}_n^k,t_{ij}\\ \mathrm{H}_n,\mathrm{R}_n^{\omega},\ldots}}\ \alpha_{\mathrm{B}}\sum_{\substack{v_n\in\mathscr{V}\\ \omega\in[1,\Omega]}}\sum\mathrm{R}_n^{\omega}+\alpha_{\mathrm{A}}\sum_{f^k\in\mathscr{F}}\mathrm{A}_{s^k}^k-\alpha_{\mathrm{N}}\sum_{\substack{v_n\in\mathscr{V}\\ f^k\in\mathscr{F}}}\sum\mathrm{A}_n^k$$

subject to:

(3.2) to (3.12), (3.16) to (3.19) and (3.22)

$$t_{ij}\in[0,1]\subset\mathbb{R},\text{all other variables}\in\{0,1\} \tag{3.23}$$

where the normalizing factors defined by:

$$\alpha_{\mathrm{B}}=\frac{1}{\Omega\,|\mathscr{V}|},\alpha_{\mathrm{A}}=\frac{1}{|\mathscr{F}|},\alpha_{\mathrm{N}}=\frac{1}{|\mathscr{V}|} \tag{3.24}$$

Algorithm 1 JRW-D2D pseudo-code

for each SL frame τ **do**
  for each $f^k \in \mathscr{F}_{\mathrm{A}}$ **do** ▷ Arriving flows
    if s_k and $d_k \in$ the same component of $\mathbb{G}$ **then**
      $\mathscr{F}_{\mathrm{W}} \leftarrow \mathscr{F}_{\mathrm{W}} \cup \{f^k\}$
    end if
  end for
  for each $f^k \in \mathscr{F}_{\mathrm{FIN}}$ **do** ▷ Finished flows
    $\mathscr{V}_{\mathrm{D}} \leftarrow \mathscr{V}_{\mathrm{D}} \cup \mathrm{NodesOF}\left(\mathbb{P}^k\right)$
  end for
  Construct the MILP model as in formula (3.23)
  Solve the MILP model using Algorithm 2
  for each $f^k \in \mathscr{F}_{\mathrm{W}}$ **do**
    if $\mathrm{A}_{s_k}^k = 1$ **then** ▷ Flow is admitted
      Configure the path according to $\mathbb{P}^k$
    end if
  end for
  $p \leftarrow \tau \mod 2$
  for each $v_n \in \mathscr{V}_{\mathrm{H}}^p$ **do**
    Allocate RBs according to $[\mathrm{R}_n^1, \mathrm{R}_n^2, \cdots, \mathrm{R}_n^{\Omega}]^{\mathrm{T}}$
  end for
end for

3.3 Proposal: JRW-D2D

In this chapter, we propose novel strategy named **Joint Routing and Wireless allocation in D2D communications** (JRW-D2D) to solve the optimization problem described above. Our proposal is based on *Branch-and-Cut* algorithm [67]. The latter is a well-known optimization algorithm and efficient to solve the general class of *Mixed-Integer-Linear-Programming* (MILP) problems. JRW-D2D proceeds as follow. First, the binary variables are relaxed by allowing them to admit continuous values between 0 and 1. Then, the relaxed problem is solved by the *simplex* algorithm. If the latter converges to an optimal solution with at least fractional value for a variable, then a *branch* is introduced on that variable. A branch means that two sub-problem nodes are scheduled to be solved

recursively with additional *cuts* (i.e., additional inequality constraints). Each cut bounds the variable in sub-problems by 0 or 1. Each sub-problem is, in its turn, relaxed again, and the whole process repeats until finding a set of feasible integral solutions that includes the optimal one. However, a scheduled problem node is *pruned* if its objective-function value in the relaxed solution is worse than the best integral solution found so far. Pruning a node means that the latter cannot generate further sub-problems. Hence, an extensive search for an optimal integral solution is avoided. Algorithms 1 and 2 illustrate the pseudo-code of our proposal `JRW-D2D`. It should be noted that we also introduce a bound on the number of recursive iterations to limit the execution time.

Algorithm 2 MILP resolution

Input: MILP Model P^0 as defined in formula (3.23)

Output: Solution value for $\mathbf{V}^*$ as $[x_{ij}, \mathrm{A}_n^k, t_{ij}, \mathrm{H}_n, \mathrm{R}_n^{\omega}, \ldots]$

1: Push the initial problem P^0 onto the stack $\mathscr{S}$
2: $f^* \leftarrow -\infty$ ▷ Initial value for Objective function
3: $\mathrm{I} \leftarrow 0$ ▷ Counter
4: **while** $\mathscr{S} \neq \emptyset \wedge \mathrm{I} \leq \mathrm{I}_{\max}$ **do**
5: $\mathrm{I} \leftarrow \mathrm{I} + 1$
6: Pop a problem from $\mathscr{S}$ as P
7: Let $\tilde{\mathrm{P}}$ be the relaxed form of P with continuous $\mathbf{V}^*$
8: Solve $\tilde{\mathrm{P}}$ using simplex yielding $\tilde{\mathbf{V}}$ and $\tilde{f}$
9: **if** not feasible or $\tilde{f} \leq f^*$ **then go to** 17
10: **if** $\tilde{\mathbf{V}}$ are all 0 or 1 except for t_{ij} **then**
11: $\mathbf{V}^* \leftarrow \tilde{\mathbf{V}}$, $f^* \leftarrow \tilde{f}$ and **go to** 17
12: **else**
13: Choose the closest variable to 0.5 as $v^{\dagger}$
14: Add a cut $v^{\dagger} \leq 0$ to P and push it onto $\mathscr{S}$
15: Add a cut $v^{\dagger} \geq 1$ to P and push it onto $\mathscr{S}$
16: **end if**
17: **end while**
18: **return** the solution value $\mathbf{V}^*$

3.4 Performance Evaluation

In this section, we will gauge the performance of our proposal JRW-D2D based on extensive simulations. First of all, we will briefly describe the network simulation environment NS-3, which we augmented to support the LTE-D2D protocol stack. Then, we will detail the studied scenario in this chapter. Afterward, we will define the performance metrics. Finally, we will analyze the simulation results and discuss the effectiveness of our proposal.

3.4.1 Network Simulation Environment

The NS-3 software package [69], which is written in C++, provides powerful open-source tools to implement a wide variety of network simulation scenarios and applications using different degrees of abstractions and reference technologies. NS-3 provides substantial support for a variety of conventional 3GPP LTE simulation scenarios through the module NS-3/LTE [70]. Unfortunately, the latter does not support the LTE-D2D standard. To the best of our knowledge, this is the case for all available network simulators in this respect. This is, in part, due to the fact that LTE-D2D is a relatively new standard. To achieve our goal, we extended the NS-3/LTE modules to include the necessary LTE-D2D protocol stack. We developed the PHY, MAC, and PDCP/RLC procedures along with the signaling between the eNB and UEs. The signaling is necessary to i) configure the SL parameters, ii) establish the SL radio bearers (SLRB), and iii) exchange the SL reports and grants.

3.4.2 Network Simulation Setup

In line with our formulation in section 3.2, we run simulations for a network composed of one macro-cell LTE-A with radius Rcell = 1 km. The geographical

deployment of UEs inside the cell follows a Poisson Point Process distribution with a density λ_{UE} nodes per km^2 for values in the set $\{10,15,20,25,30,35,40\}$. The LTE-A macro-cell is configured to work in FDD mode with a UL frequency of 1930 MHz and a bandwidth of 5 MHz (i.e., 25 RBs). The eNB configures SL bandwidth to share the same as UL. However, The eNB allocates scheduled-resources pool only $\Omega = 14$ RBs for the offloading operation over SL. UEs are configured to transmit on SL with a common power density of $\Psi_t = -4$dBm/RB (i.e., maximum of 10 dBm over the whole 5 MHz). To model the SL path-loss, we use the WINNER II B2-LOS channel model [71]. The SL-Period (SL frame) is configured to be 40 milliseconds (i.e., 40 subframes), which is the minimum possible value in the standard, of which 32 subframes are used for the data transmission. The eNB, using SNR reports, builds the D2D network topology. A link is considered part of the network if the respective SNR is greater than $\rho_{TOPO} = 10$ dB. Traffic flows are generated according to a Poisson process with an arrival rate of $\lambda_{FL} \in \{10,20\}$ flows per second. On the other hand, each flow is assumed to have a Constant Bit-Rate (CBR) traffic randomly selected from pre-defined CBR classes. Flow duration distribution is simulated to follow an exponential random variable with a mean duration of $\lambda_{DUR} = 1$ second. Sources and destinations are chosen from a random uniform distribution. Table 3.1 summarizes the main parameters used in our network simulation. For the evaluation of results, the confidence level is set to 95%.

3.4.3 Performance Metrics

Let $\mathscr{F}_{TOT} \supseteq \mathscr{F}_{ADM}$ be the total sets of arrived, and the admitted flows, respectively, during a simulation run. Also, let $E[\cdot]$ denote the average sample metric over all the simulation runs. Then, We define the following metrics to evaluate our proposal:

Table 3.1 – Simulation Parameters

Parameter	**Value**
Cell Radius Rcell	1 km
UL/SL Frequency f_{UL}	1930 MHz
UL/SL (Reference) Bandwidth B_{UL}	5 MHz (25 LTE RBs)
SL RBs Used Actually Ω	14 LTE RBs
SL frame (LTE-D2D SC-Period)	40 subframes (40 ms)
Data Part in SL frame	32 subframes
UE SL Power Transmit Density Ψ_t	-4 dBm/RB
Noise Spectral Density Ψ_n	-121.45 dBm/RB
LTE MCS Index used in SL	9 (QPSK)
UE Density λ_{UE}	$\{10,15,20,25,30,35,40\}$ nodes per km^2
UE-UE SNR Threshold γ_{TOPO}	10 dB
Scheduling SINR Threshold γ	6 dB
Flow Simulation Period	10 seconds
Flow Arrival Process	Poisson Process
Flow Arrival Rates λ_{FL}	$\{10,20\}$ flows/second
Flow Duration Random Variable	Exponential
Flow Duration Mean λ_{DUR}	1 second
Flow Bit Rate Classes	$\{25,50,75,100,125,150,175,200\}$ kbps

1. $\mathbb{S}$ is the ratio of the flows offloaded by the D2D network. This metric $\mathbb{S}$ is defined by:

$$\mathbb{S} = \mathrm{E}\left[\frac{|\mathscr{F}_{\mathrm{ADM}}|}{|\mathscr{F}_{\mathrm{TOT}}|}\right] \tag{3.25}$$

2. $\mathbb{A}$ is the maximum number of scheduled flows simultaneously. Formally, this metric $\mathbb{A}$ is defined as:

$$\mathbb{A} = \mathrm{E}\left[\max_{\tau} |\mathscr{F}_{\mathrm{ADM}}[\tau]|\right] \tag{3.26}$$

3. $\mathbb{H}$ is the average number of hops in the offloading path in each simulation

run. Formally, this metric $\mathbb{H}$ is defined as:

$$\mathbb{H} = \mathrm{E}\left[\frac{\sum_{f^k \in \mathscr{F}_{\mathrm{ADM}}} \textsc{HopsOF}\left(\mathbb{P}^k\right)}{|\mathscr{F}_{\mathrm{ADM}}|}\right] \tag{3.27}$$

4. $\mathbb{L}$ is the average of flow packet loss in each simulation run. Formally, this metric $\mathbb{L}$ is defined as:

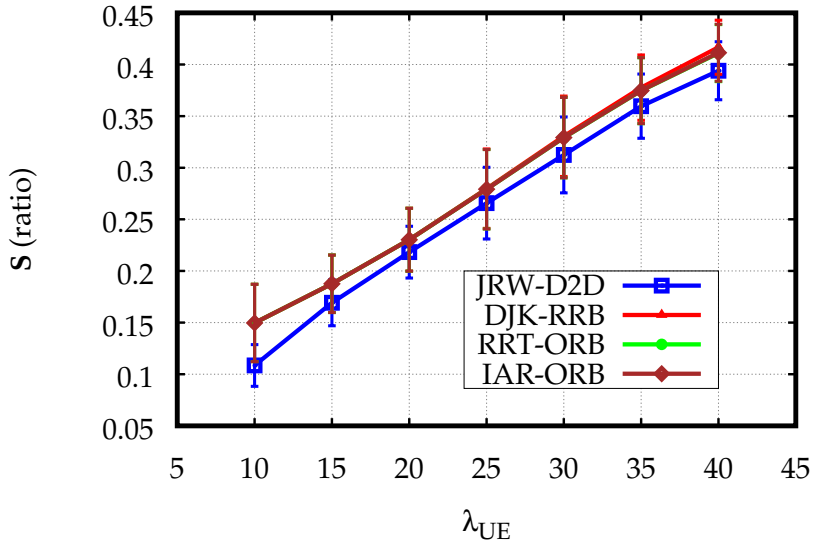
$$\mathbb{L} = \mathrm{E}\left[\frac{\sum_{f^k \in \mathscr{F}_{\mathrm{ADM}} \setminus \mathscr{F}_{\mathrm{INT}}} \frac{\mathtt{pkts}^k_{\mathrm{tx}} - \mathtt{pkts}^k_{\mathrm{rx}}}{\mathtt{pkts}^k_{\mathrm{tx}}}}{|\mathscr{F}_{\mathrm{ADM}}|}\right] \tag{3.28}$$

Moreover, we compare the performance of our proposal `JRW-D2D` with the following alternative routing and OFDMA resource allocation strategies:

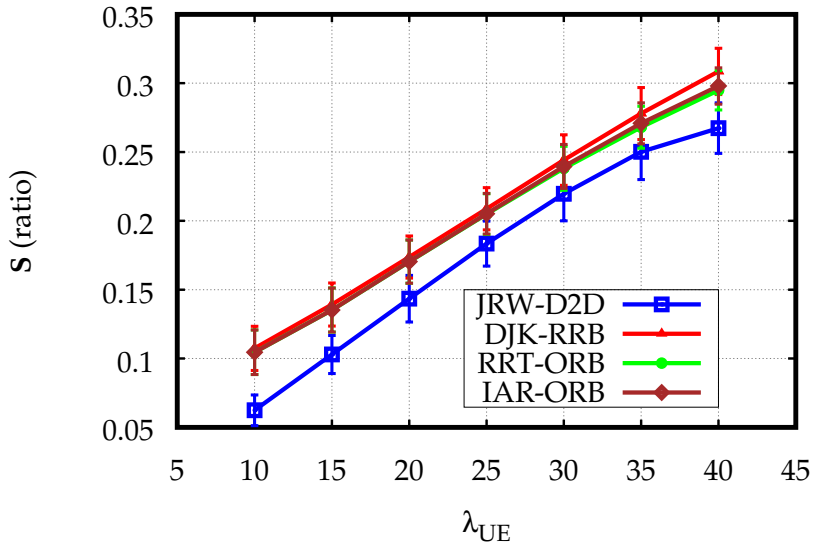
1. `DJK-RRB`: is a pure path strategy that aims to find the optimal routing trees using the Dijkstra algorithm and then allocates RB randomly.

2. `RRT-ORB`: is a pure resource block-oriented strategy that finds the routing trees randomly using random walk on the topology graph, and allocates RB optimally.

3. `IAR-ORB`: is a heuristic scheme composed of interference aware routing based on the Dijkstra algorithm. In this variant, the link costs to minimize are the total interference level on the link taking into consideration the actual state of the network before accepting the new flows. Then, the resource block allocations are done optimally.



(a) $\lambda_{FL} = 10$



(b) $\lambda_{FL} = 20$

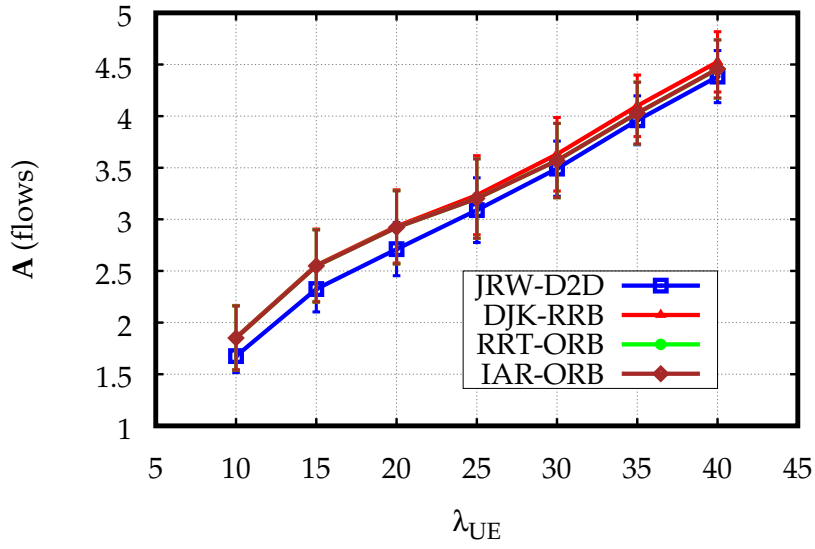
Figure 3.2 – $\mathbb{S}$ versus nodes density λ_{UE}.

3.4.4 Simulation Results

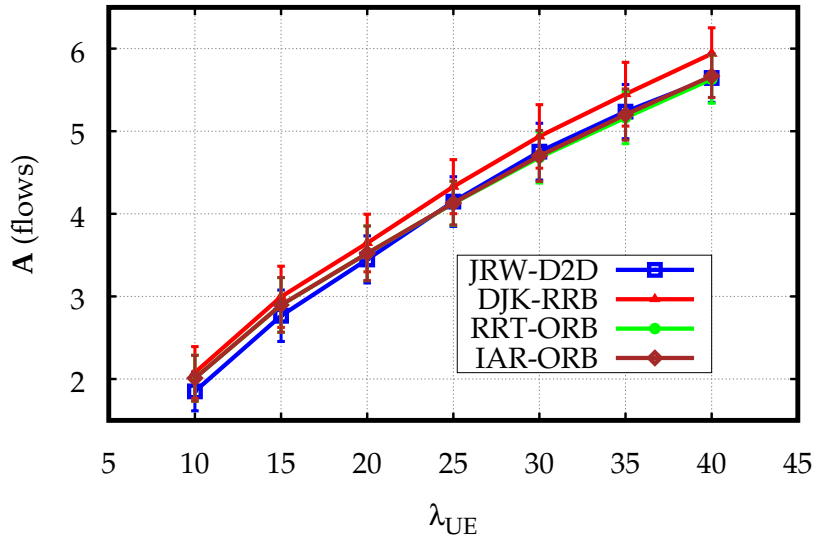
First, we evaluate our algorithm `JRW-D2D` regarding its offloading capability and in comparison with the alternative strategies `DJK-RRB`, `RRT-ORB`, and `IAR-RRB`.

To do so, we calculate the ratio of the flows offloaded over the D2D network. Figure 3.2 illustrates $\mathbb{S}$ with respect to the density of UEs and under two traffic conditions $\lambda_{FL} = 10$ and $\lambda_{FL} = 20$ flows per second. For small values of λ_{UE}, it is straightforward to see that $\mathbb{S}$ increases in proportion to λ_{UE}. This means that as D2D node density increases, more flows will succeed to be routed through the D2D network. This is expected because, when the density increases, the probability of forming reliable D2D links rises accordingly. And as a consequence, the D2D network capacity to absorb random flows also grows. We remark that `DJK-RRB` outperforms the other schemes in general. This is expected since `DJK-RRB` routes the flows over the fewest possible nodes. As a result, it allows for more flows to be admitted into the network. Taking `DJK-RRB` as a baseline, we note that our proposal `JRW-D2D` has a flow acceptance rate of $\Delta\mathbb{S} = 1\%$ less than the baseline `DJK-RRB`, in average, for $\lambda_{FL} = 10$ as depicted in Figure 3.2 (a). On the other hand, Figure 3.2 (b) shows the situation under more traffic pressure, $\lambda_{FL} = 10$, where the acceptance rate drops for all schemes while the performance gap of `JRW-D2D` increases to be around $\Delta\mathbb{S} = 3\%$ in average with respect the leader `DJK-RRB`.

To complement the evaluation of the offloading capability, we measure the degree concurrency in utilizing the D2D network. To this end, we measure the average of the maximum number of flows offloaded simultaneously over the D2D network. This measure is conveyed by the metric $\mathbb{A}$, shown in Figure 3.3, which demonstrates to what degree the different schemes are successful in utilizing system resources concurrently. In a manner consistent the evolution of $\mathbb{S}$, the evolution of $\mathbb{A}$ is depicted in Figure 3.3 under the two traffic conditions $\lambda_{FL} = 10$ and $\lambda_{FL} = 20$. We note that the metric $\mathbb{A}$ increases in response to an increase in λ_{UE}. This reflects the fact that, in a denser topology, more nodes are available in the network to route concurrent flows circumventing the restric-



(a) $\lambda_{FL} = 10$

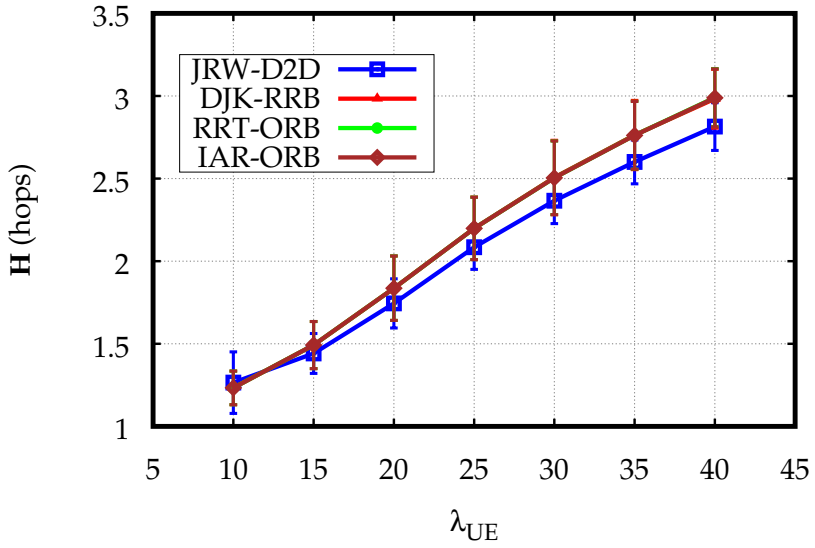


(b) $\lambda_{FL} = 20$

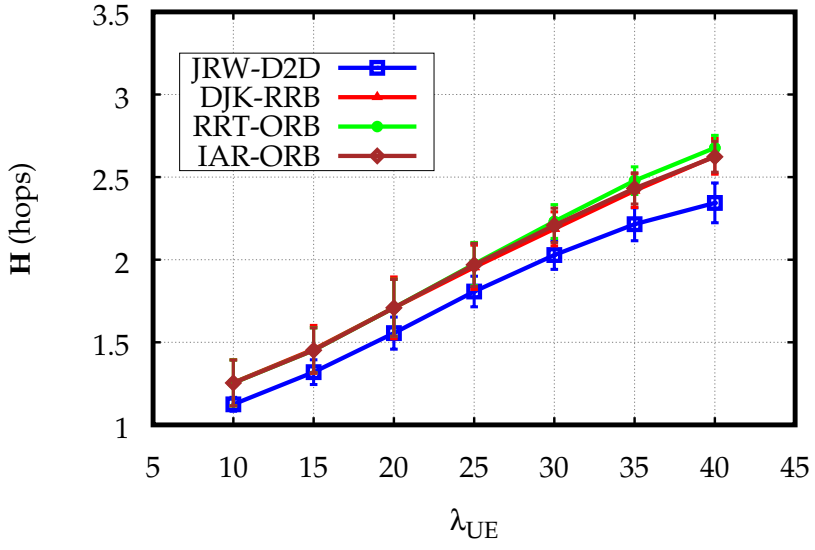
Figure 3.3 – $\mathbb{A}$ versus node density λ_{UE}.

tion due to maximum one flow per node. Again, we note that `DJK-RRB` is the leader of the group where the our scheme `JRW-D2D` was able to offload slightly

fewer simultaneous flows than the others with performances gaps $\Delta\mathbb{A} \leq 0.25$ simultaneous flows as indicated in Figure 3.3 (a) and Figure 3.3 (b).



(a) $\lambda_{FL} = 10$



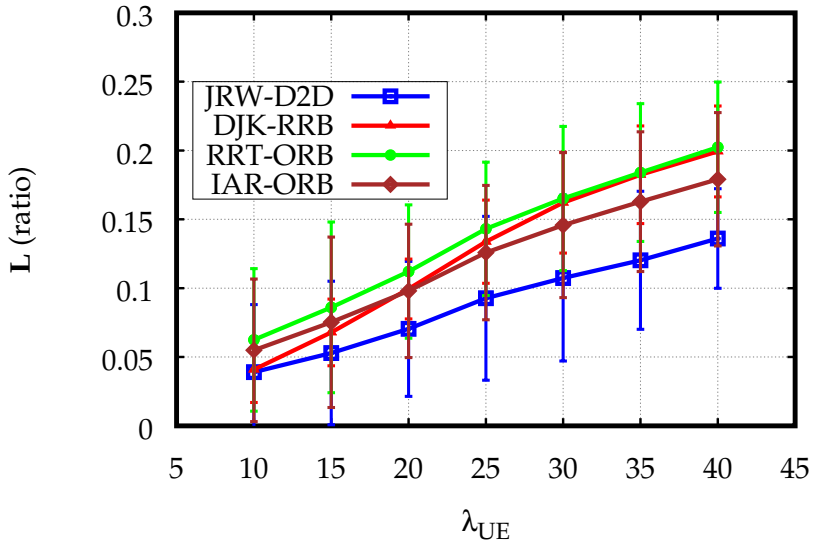
(b) $\lambda_{FL} = 20$

Figure 3.4 – $\mathbb{H}$ versus node density λ_{UE}.

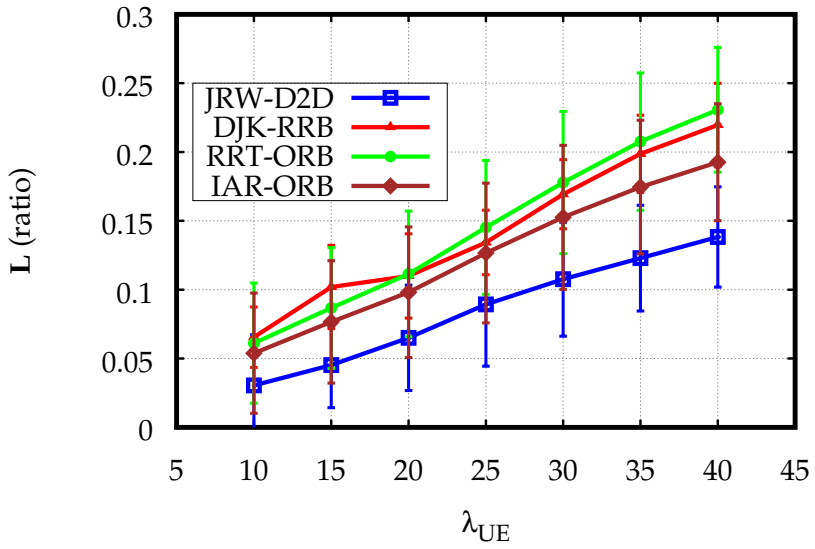
To quantify the QoS presented to the offloaded flows we focus on latency and packet loss rate. Figure 3.4 illustrates the performance in terms of $\mathbb{H}$ metric, which count the number of hops in the routing paths. This metric indicates the QoS presented to flows in terms of latencies where shorter is better. Specifically, the end-to-end and the average packet delays are proportional to $\mathbb{H} \times T_{SL}$. Figure 3.4 (a) and Figure 3.4 (b) point out that the average number of hops increases almost linearly in accordance with the density of nodes λ_{UE}. The figure reveals that JRW-D2D leads to shorter paths on average than the others. This seems paradoxical in particular when comparing to DJK-RRB. However, lower values of $\mathbb{H}$ are an artifact of JRW-D2D being biased to accept flows with shorter paths at the expense of blocking some long path flows.

Moreover, to quantify the QoS in terms of the packet error rate at the IP level, Figure 3.5 illustrates the average packet loss ($\mathbb{L}$) in flows as a function of the UEs' density for $\lambda_{FL} = 10$ and $\lambda_{FL} = 20$ conditions respectively. In Figure 3.5, it is straightforward to see that our scheme JRW-D2D outperforms the other schemes thanks to their capability to take into consideration interference in OFDMA RB blocks allocation. However, RRT-ORB performs poorly in general, which may seem paradoxical. It is straightforward to see that such behavior will lead to higher transmission delays. Being interference-aware in routing and resource allocation makes JRW-D2D more robust against packet loss. In fact, the latter succeeds to maintain $\mathbb{L}$ below 0.13 and 0.14 for both traffic conditions $\lambda_{FL} = 10$ and $\lambda_{FL} = 20$ flow per seconds respectively as depicted in Figure 3.5 (a) and Figure 3.5 (b).

In summary, network simulations show that JRW-D2D outperforms the variants in terms of reliability at the expense of small performance gaps with respect to latency and offloading capacity.



(a) $\lambda_{FL} = 10$



(b) $\lambda_{FL} = 20$

Figure 3.5 – $\mathbb{L}$ versus node density λ_{UE}.

3.5 Conclusion

In this chapter, we addressed the problem of joint routing and OFDMA resource allocation in LTE-D2D multihop networks, considering LTE-D2D-specific con-

straints, namely, the half-duplex operation and the contiguity in RB allocations. We presented an offloading application use case where data from unicast flows are routed over the D2D multihop network, and the eNB hosts the control plane. To optimize offloading, we proposed a MILP formulation for the problem and a novel scheme named JRW-D2D based on the branch-and-cut algorithm. Next, we validated our proposal by simulating the whole LTE D2D protocol stack in the NS-3 network simulator. We compared our JRW-D2D to other basic one-sided optimal strategies and another interference-aware heuristic scheme. The results obtained are very satisfying in terms of optimality, the ratio of admitted D2D flows, and latency.

Potential enhancements to JRW-D2D include extending its scope to consider multicast UE-to-UEs flows in the cell. Besides, we must evaluate the scalability of JRW-D2D in dense D2D topologies. As shown next, schemes based on ILP, such as JRW-D2D, do not scale up easily due to the complexity of ILP models, the joint treatment, and the batch processing of the incoming flows.

Chapter 4

A Scalable Joint Routing and Resource Allocation Scheme: D2D-based Unicast and Multicast Data Offloading

Contents

5G networks take advantage of a wide range of novel technologies to respond to the massive traffic workloads of the recent Zettabyte era. In this context, Device-to-Device (D2D) communications can bring solutions to the cellular network's problems as regards congestion, power consumption, and effi-

cient use of the frequency spectrum. In this chapter, we address the optimal design of scalable offloading schemes based on the LTE-D2D standard to offload intracellular traffic, unicast, and multicast, over multihop D2D networks of cooperative User-Equipments. Specifically, we deal with the problem of joint routing and OFDMA resource allocation that underlies such schemes. Considering crowded-platform use-cases [72], we propose a novel path-based ILP formulation in which a routing tree is formulated in terms of its constituent paths. Moreover, to boost scalability, we propose a sub-optimal solution method, named `JRW-D2D-CG`, based on the column-generation framework with a pricing problem. The latter allows us to consider only paths that are likely to enhance the solution. We adjust the pricing problem to be more tractable, and then, we use a fast algorithm based on the Bellman-Ford algorithm to find advantageous paths. Based on extensive network simulation in the NS-3 environment, we show that our proposal achieves good performances in terms of reliability, latency, and scalability.

4.1 Introduction

In the previous chapter, we tackled the problem of joint routing and OFDMA resource block allocation that lies beneath the management of a multihop D2D offloading scheme for the intracellular traffic. In this chapter, we propose an offloading scheme based on LTE-D2D to route intracellular UE-to-UE, both unicast and multicast traffic, through a network of the D2D UEs.

In the same vein as the previous chapter, the eNB aims to offload the infrastructure from the traffic exchanged between the UEs within its cellular area where the targeted setup consists of an LTE-A FDD omnidirectional macro-cell. We assume a topology of quasi-stationary D2D-connected UEs which work to-

gether, under the supervision of the eNB, to carry the traffic data flows through multiple hops to reach destinations. Under the quasi-stationarity assumption, the relaying service presents always-on connections to flow-centric applications. However, to guarantee such a service, the eNB takes charge of i) computing routes over the D2D topology, ii) allocating enough RBs to the relay nodes, and iii) managing interferences. Use cases of such network service include crowded-platform scenarios, for example, user terminals in stadiums or the waiting halls in airports and train stations.

Regarding the reviewed literature, unlike [44], [45], and, in particular, [43], we select a semi-static routing where the path computation considers the current state of the interfering links. Nevertheless, our approach keeps each path remains unchanged for the entire period of communication to deliver a continuous always-on relaying operation. We also model the spectrum allocation to the resource-block level and consider the fact that LTE-D2D allocates them contiguously, which is a 3GPP constraint for both UL and SL. Unlike [52], we consider delay-sensitive traffic. Also, as opposed to [54] and [50], we make use of the standardized D2D links to offload the multicast traffic that starts and ends in the UEs themselves. Compared to [53], our scheme extends the relaying operation to the multihop level. We address the problems of routing and resource allocation that underlie the offloading scheme. We formulate the joint problem as an Integer Linear Programming (ILP) model, which considers spectrum reuse constraints as well as other LTE-D2D limitations: half-duplex operation and contiguity of resource block allocations.

For the efficient use of this offloading scheme, the eNB must optimize the routing and resource allocation. Precisely, the eNB has to build some optimization model, which also considers the specific LTE-D2D constraints, such as: the contiguity of the RBs and the half-duplex transmission mentioned above.

Besides, the eNB must solve the optimization model quickly, and the solution time should scale well with the number of UEs in the system to handle dense scenarios.

We summarize the most important contributions of this chapter as follows:

i) We introduce two Integer Linear Programming (ILP) formulations for the problem. The first is a link-based formulation to be solved using the well-known Branch-and-Cut method. The second formulation is a novel path-based one, which is, albeit extensive (i.e., has a huge number of variables), can be solved efficiently using the column-generation approaches.

ii) We present a new graph-based formulation for the OFDMA RB allocation that leads to a more tractable optimization model than the matrix-based formulation introduced in [40].

iii) More importantly, we propose a sub-optimal quick solution method, named `JRW-D2D-CG`, to solve the path-based ILP model using a column-generation framework in which we iteratively add "good" paths to the main problem using a fast sub-optimal procedure.

iv) Through extensive network simulations in NS-3, which we extended to support LTE-D2D, we evaluate the performance of our proposal as compared to the one introduced in [40], named `JRW-D2D-NS`, and we also include a scheme based on Dijkstra's shortest-path algorithm.

Simulation results show that our proposal `JRW-D2D-CG` achieves comparable performance to `JRW-D2D-NS` in terms of reliability and latency. However, our new proposal `JRW-D2D-CG` scales much better with the size of the D2D topology when compared to the previous `JRW-D2D-NS` which takes so long to solve the underlying ILP model to optimality.

The rest of this chapter is organized as follows. In Section 4.2, we detail, in-depth, the description of our model and two formulations that underlie the problem mentioned earlier. Moreover, Section 4.3 presents our proposed scheme to solve the problem leveraging the extensive path-based formulation. In Section 4.4, we discuss the network simulation results and a comparative evaluation of our proposal relative to the related schemes. Lastly, Section 4.5 concludes the chapter summarizing the main results.

4.2 System Model and Problem Formulations

In this section, we describe two formulations for our system model: the initial link-based formulation [40] and the new one, introduced here, based on the **column-generation** method. We suppose that we have N UEs inside the zone of a cell controlled by one LTE-A eNB equipped with an omnidirectional antenna. Additionally, we assume that the UEs produce one-to-many traffic flows from and towards the UEs themselves. Conventionally, this intracellular traffic must pass through the eNB (and the core network). However, since our UEs support the LTE-D2D, we assume that they can deliver the traffic directly by themselves using multihop D2D-links (i.e., SL) but still under the control of the eNB where resides a central algorithm that optimizes this offloading operation. We also assume that the topology of the D2D network is quasi-static, and therefore it does not undergo short-term disruptions.

This assumption reduces the signaling required to monitor the wireless links qualities and to do the buffer management, and therefore it facilitates system design. Nevertheless, this does not limit the practicality of our design since many crowded-platform scenarios meet this condition, such as content-sharing applications in stadiums and the waiting-halls in airports and train stations. In

such scenarios, the UEs are stationary most of the time, and hence the assumption of nodes stationarity is justified.

The role of the eNB is to find, if possible, a routing tree (or path) for each flow arriving in its waiting queue. In line with the RB allocation, we suppose that the routing decisions have the same time scale. In other words, the eNB decides on routing and RB allocation in every SL frame, at the decision instants $t = \tau \times T_{SL}$ where $\tau \in \mathbb{N}$ is the index of the SL frame.

To be consistent with a flow-centric offloading, we design the relaying process to present an always-on connection. To this end, we assume that the UEs relays (i.e., the nodes) are exclusive: once the eNB allocates a routing tree (or path) for a flow, it keeps it for the entire duration of the flow. However, note that the RB allocations of the active nodes are subject to change in response to the network load (i.e., the departure, and arrival of flows) and the radio environment (i.e., interferences, and noise).

To keep the always-on connected-mode relaying, one must deal with the half-duplex mode restriction in LTE-D2D. Consequently, we employ an alternating link-activation approach. In this approach, the eNB activates the links that are one-hop apart during the same SL frame while those in-between are idle, and then it reverses the situation in the next SL frame. Stated in terms of nodes, active nodes transmit every other SL frame while their direct receivers listen to them, and then they switch roles in the next SL frame.

To model the D2D topology, similar to Chapter 3, we use a (symmetric) directed graph $\mathbb{G} = (\mathcal{V}, \mathcal{E})$. The set $\mathcal{V} = \{v_n \mid n = 1, 2, \cdots, N\}$ denotes the set of the N UEs (i.e., the nodes) and $\mathcal{E}$ denotes the set of viable communication links between the UEs. A link e_{ij} between the nodes v_i and v_j is considered to be a part of the topology (i.e., $e_{ij} \in \mathcal{E}$) if it acheives, at least, a given targeted Signal-to-

Noise Ratio (SNR). More formally,

$$\gamma_{ij} = \frac{g_{ij} \mathrm{P}_{t,i}}{\mathrm{P}_{\sigma}} \geq \gamma_{\mathrm{TOPO}} \iff e_{ij} \in \mathscr{E} \tag{4.1}$$

where i) γ_{ij} denotes the SNR between v_i and v_j, ii) $\mathrm{P}_{t,i}$ denotes the power emitted from the UE v_i, iii) P_{σ} denotes the noise power, iv) g_{ij} represents the channel gain between v_i and v_j which depends on the channel model to use, and v) γ_{TOPO} is threshold value set according to a targeted performance. Note that the eNB constructs this topology model by collecting *long-term* periodic measurements of the quality of radio links in the network.

With regard to the traffic model, we adopt a unified one-to-many model, where each flow f_k, with index k, has a source node $s^k \in \mathscr{V}$ and a set of destination nodes $\mathscr{D}^k \subseteq \mathscr{V}$. In this model, we consider both multicast and unicast applications uniformly since *one can consider a unicast flow as a particular case where the destination set includes a single node* (i.e., $\left|\mathscr{D}^k\right| = 1$). Note also that for the unicast case, the required routing tree reduces to a simple path.

We also assume that our flows are generated by Constant Bit-Rate type (CBR) applications, as in Chapter 3, where the bit-rate of the flow R^k is mapped directly to a requested number of RBs, RB^k using the equation (3.13).

The proposed offloading scheme strives to increase the utility of the SL's spectrum by reusing the available OFDMA RBs. To increase efficiency, it can reallocate the same RB to several nodes if it can keep the mutual interference below a harmful level. In this regard, we applied the same per-RB treatment outlined in Section 3.2 assuming the same fixed emission power density Ψ_t [mW/RB], for all the nodes, and the same flat block-fading channel model. Again, the eNB verifies that the SINR $\tilde{\gamma}_{ij}$ of each active link is under a speci-

fied level γ which can be written as follows:

$$\tilde{\gamma}_{ij} \leq \gamma \iff \sum_{v_n \in \mathcal{V}_{\text{interfering}}} g_{nj} \Psi_{t,n} + \Psi_{\sigma} \leq \frac{g_{ij} \Psi_{t,i}}{\gamma} \tag{4.2}$$

We summarize the task of the eNB in every SL frame, with time index τ, in the following steps:

UPDATE STEP: The eNB updates the topology model in response to changes in the radio environment and flow-exit events.

ADMIT STEP: The eNB evaluates the flows in the waiting queue $\mathscr{F}_{\text{WAIT}}$ to accept some into the offloading system.

ROUTE STEP: A flow f^k is only accepted if:

1. There is a dedicated routing tree $\mathbb{T}^k$ over the passive nodes.
2. There are valid half-duplex set assignments for the nodes in $\mathbb{T}^k$.
3. There are valid corresponding RB assignments such that the whole system avoids harmful interferences.
4. Once the eNB accepts some flow f^k, it holds the routing tree and the respective half-duplex assignments until the flow finishes.

RB ALLOCATION STEP: The eNB recomputes the RB allocations for all the flows in progress $\mathscr{F}_{\text{SCHED}}$ while considering the interferences.

GRANT STEP: The eNB sends the RB allocations that match the current half-duplex set $\mathcal{V}_{\text{H}}^{q}$, where $q = \tau \mod 2$.

We propose to solve the routing step while considering the subsequent RB allocation step, the wireless environment, and the available RBs in the system. Also, we propose to evaluate all the waiting flows for admission at the same time

instead of one-by-one evaluation. To sum up, we propose a joint treatment for the routing and resource allocation that underlie a bulk-queue offloading service presented by our D2D network to relieve the cellular infrastructure from the burden of carrying the whole traffic alone.

Note that the goal of the eNB is to admit (i.e., to offload) the maximum number of the flows into the system. When doing so, it endeavors to satisfy their bit-rate demand, so that they finish as early as possible while being parsimonious in resources, so no relay nodes or RBs are used beyond what is needed.

4.2.1 Initial Link-Based Formulation

In the following paragraphs, we describe the initial formulation for the routing and resource block allocation described so far. In this formulation, denoted by `JRW-D2D-NS`, we describe the routing side of the decision problem using per-link per-level 0-1 variables to encode routing trees. To route some flow $f^k \in \mathscr{F} = \mathscr{F}_{\text{WAIT}} \cup \mathscr{F}_{\text{SCHED}}$, we solve for 0-1 variables $x_{ij}^{h,k}$ which mirror the links $\mathscr{E}$. A variable $x_{ij}^{h,k}$ determines whether the respective link e_{ij} is in the tree $\mathbb{T}^k$ at the hop number h (i.e., tree level) for $h = 0, 1, 2, \cdots, h_{\max}$. However, to make sure that the link selection is compatible with the mathematical definition of a tree and compatible the node-exclusivity mentioned above, we use additional auxiliary variables to formulate the routing constraints as linear inequalities as follows:

$$\sum_{\substack{e_{ij}\in\mathscr{T}(v_n)\\ 0\le h\le h_{\max}, f^k\in\mathscr{F}}}\sum\sum x_{ij}^{h,k} \le 1 \qquad \forall v_n\in\mathscr{V} \tag{4.3}$$

$$x_{ij}^{h,k} \le \delta_0^h \cdot \delta_{v_i}^{s^k} + (1-\delta_0^h)(1-\delta_{v_i}^{s^k}-\delta_{v_j}^{s^k}) \qquad \begin{matrix}\forall e_{ij}\in\mathscr{E}\\ \forall 0\le h\le h_{\max}\\ \forall f^k\in\mathscr{F}\end{matrix} \tag{4.4}$$

$$x_{nm}^{h,k} \le \sum_{e_{ij}\in\mathscr{T}(v_n)} x_{ij}^{h-1,k} \qquad \begin{matrix}\forall e_{nm}\in\mathscr{E}\\ \forall 1\le h\le h_{\max}\\ \forall f^k\in\mathscr{F}\end{matrix} \tag{4.5}$$

$$\sum_{\substack{e_{ij}\in\mathcal{T}(v_n)\\0\le h\le h_{\max}}}\sum x_{ij}^{h,k}-\sum_{\substack{e_{ij}\in\mathcal{O}(v_n)\\0\le h\le h_{\max}}}\sum x_{ij}^{h,k}\le \mathbb{1}_{v_n}^{\mathcal{D}^k} \qquad \begin{matrix}\forall v_n\in\mathcal{V}\\ \forall f^k\in\mathcal{F}\end{matrix} \tag{4.6}$$

$$t_n^k \ge \sum_{0\le h\le h_{\max}} x_{nm}^{h,k} \qquad \begin{matrix}\forall e_{nm}\in\mathcal{E}\\ \forall f^k\in\mathcal{F}\end{matrix} \tag{4.7}$$

$$t_n^k \le \sum_{\substack{e_{ij}\in\mathcal{O}(v_n)\\0\le h\le h_{\max}}}\sum x_{ij}^{h,k} \qquad \begin{matrix}\forall v_n\in\mathcal{V}\\ \forall f^k\in\mathcal{F}\end{matrix} \tag{4.8}$$

$$\sum_{\substack{e_{ij}\in\mathcal{T}(v_n)\\0\le h\le h_{\max}}}\sum x_{ij}^{h,k} \ge \mathbb{1}_{v_n}^{\mathcal{D}^k}\cdot t_{s^k}^k \qquad \begin{matrix}\forall v_n\in\mathcal{V}\\ \forall f^k\in\mathcal{F}\end{matrix} \tag{4.9}$$

$$\sum_{f^k\in\mathcal{F}} t_n^k \le 1 \qquad \forall v_n\in\mathcal{V} \tag{4.10}$$

$$\sum_{f^k\in\mathcal{F}}\left(\delta_{v_n}^{s^k}+\mathbb{1}_{v_n}^{\mathcal{D}^k}\right)\cdot t_{s^k}^k \le 1 \qquad \forall v_n\in\mathcal{V} \tag{4.11}$$

where we introduce the auxiliary variable t_n^k to determine when we use v_n as a sender in the tree $\mathbb{T}^k$ which depends on the variables $x_{ij}^{h,k}$'s as specified by Constraints (4.7) and (4.8). We also use the symbols $\mathcal{O}(v_n)$ and $\mathcal{T}(v_n)$ to denote the sets of outgoing edges from and incoming edges to v_n respectively. Note also that we compress the formulation using the Kronecker delta function δ_x^y, which equals to 0 unless the two arguments are equal where it takes the value 1. For the same purpose, we use the indicator function of the set $\mathcal{D}^k$ which is equal to 0 unless $v_n\in\mathcal{D}^k$ where it takes the value 1.

Constraints (4.3), (4.4), and (4.5) are related to the tree structure. Constraint (4.3) means that a node can have at most one parent node (i.e., at most, there is one incoming link selected at v_n in, at most, one routing tree. Constraint (4.4) makes sure that we can select the links from a source node only at the root of the tree (i.e., where $h=0$) and no other links can be selected at this level $h=0$. Constraint (4.5) indicates that we can select a link at the (hop) level h of a tree, only if it has a parent link at the previous level $h-1$.

Constraint (4.6) guarantees that only a destination node can be a leaf in the tree. This condition rules out trees without useless branches that do not reach

a destination. Note that a destination node can function as a relay for the same flow. Constraint (4.9) specifies that we build the routing tree for a given flow only if it is possible to reach all the destinations.

For the assignment of half-duplex sets discussed earlier, we use 0-1 variables H_n that assign the active nodes v_n to the corresponding set from $\mathcal{V}_{\mathrm{H}}^0$ and $\mathcal{V}_{\mathrm{H}}^1$. So, in this formulation, to implement the alternating link activation described above, we use the following linear constraint:

$$\sum_{\substack{f^k \in \mathscr{F} \\ 0 \le h \le h_{\max}}} \sum x_{ij}^{h,k} \le \mathrm{H}_i + \mathrm{H}_j \le 2 - \sum_{\substack{f^k \in \mathscr{F} \\ 0 \le h \le h_{\max}}} \sum x_{ij}^{h,k} \quad \forall e_{ij} \in \mathscr{E} \tag{4.12}$$

Now for the RB allocation part of the problem, we denote the number of RBs by W (i.e., $\mathrm{B_{SL}} = \mathrm{W}$ RBs). We assign to each node v_n, a vector of 0-1 variables r_n^w for $w = 1, 2, \cdots, \mathrm{W}$. The variables r_n^w indicate whether the system allocates the RB w to v_n. To enforce the contiguity restriction mentioned earlier, we represent all possible allocation vectors as columns in a matrix $\mathcal{Z}_{\mathrm{W}\times\mathrm{U}} = [z_{w,\mathrm{U}}]$. The number of columns in Z is $\mathrm{U} = \frac{\mathrm{W}(\mathrm{W}+1)}{2}$. To explain this idea, we give a simple example for the case $\mathrm{W} = 4$ as follows:

$$\mathcal{Z}_{4\times 10} = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}$$

Here the columns of $\mathcal{Z}_{4\times 10}$ enumerate all possible contiguous allocation of 4 RBs, e.g., the seventh column represents a possible case of two allocated RBs: namely the third and fourth ones.

To determine the RBs allocated to a node v_n, we use another set of 0-1 variables $y_{u,n}$ to decide which column of the matrix $\mathcal{Z}_{\mathrm{W}\times\mathrm{U}}$ encodes the allocation.

The following linear constraint relates the variables r_n^w to $y_{u,n}$:

$$r_n^w = \sum_{u=1}^{\mathrm{U}} y_{u,n} z_{w,u} \qquad \begin{matrix} \forall v_n \in \mathcal{V} \\ \forall 1 \le w \le \mathrm{W} \end{matrix} \tag{4.13}$$

To complete the formulation, we also need additional auxiliary variables derived from the previous ones: i) c_{ij}: 0-1 variable that shows if some routing tree uses e_{ij}, ii) b_n: integer variable that shows how many RBs v_n uses, iii) $r_n^{w,q}$ and $r_{ij}^{w,q}$: 0-1 variables that show if v_n and e_{ij} respectively use the RB, with index w, during the SL frames of the half-duplex set $\mathcal{V}_{\mathrm{H}}^q$, iv) $\Phi_{n,ij}^{w,q}$: 0-1 variable that shows if v_n interferes with e_{ij} on the RB, with index w, and they are both active during the SL frames of the half-duplex set $\mathcal{V}_{\mathrm{H}}^q$. The following constraints set the previous variables to their definitions:

$$c_{ij} = \sum_{\substack{0 \le h \le h_{\max} \\ f^k \in \mathcal{F}}} \sum x_{ij}^{h,k} \qquad \forall e_{ij} \in \mathcal{E} \tag{4.14}$$

$$b_n = \sum_{w=1}^{\mathrm{W}} r_n^w \qquad \forall v_n \in \mathcal{V} \tag{4.15}$$

$$r_n^{w,0} = r_n^w - r_n^{w,1}; \quad r_n^{w,1} = \mathrm{H}_n \cdot r_n^w \qquad \begin{matrix} \forall v_n \in \mathcal{V} \\ \forall 1 \le w \le \mathrm{W} \end{matrix} \tag{4.16}$$

$$r_{ij}^{w,q} = r_i^{w,q} \cdot c_{ij} \qquad \begin{matrix} \forall e_{ij} \in \mathcal{E} \\ \forall 1 \le w \le \mathrm{W} \\ \forall q \in \{0,1\} \end{matrix} \tag{4.17}$$

$$\Phi_{n,ij}^{w,q} = r_n^{w,q} \cdot r_{ij}^{w,q} \qquad \begin{matrix} \forall v_n \in \mathcal{V}, \forall e_{ij} \in \mathcal{E} \\ \forall 1 \le w \le \mathrm{W} \\ \forall q \in \{0,1\} \end{matrix} \tag{4.18}$$

Note that the products of the binary variables are linearized using a typical manner by introducing, for each term $x \cdot y$, an additional 0-1 variable λ_{xy}, and three linear constraints as follows:

$$\lambda_{xy} \le x\,(a),\ \lambda_{xy} \le y\,(b),\ \lambda_{xy} \ge x + y - 1\,(c) \tag{4.19}$$

To formulate the remaining restrictions on the resource block allocation, we

add the following constraints:

$$\sum_{u=1}^{\mathrm{U}} y_{u,n} \leq \sum_{f^k \in \mathscr{F}} t_n^k \qquad \forall v_n \in \mathscr{V} \tag{4.20}$$

$$b_n \leq \mathrm{W} + (\mathrm{RB}^k - \mathrm{W}) \cdot t_n^k \qquad \begin{matrix} \forall v_n \in \mathscr{V} \\ \forall f^k \in \mathscr{F} \end{matrix} \tag{4.21}$$

$$\Psi_\sigma r_{ij}^{w,q} + \sum_{n \neq i} g_{nj} \Psi_t \cdot \phi_{n,ij}^{w,q} \leq \frac{g_{ij} \Psi_t}{\gamma} r_{ij}^{w,q} \qquad \begin{matrix} \forall e_{ij} \in \mathscr{E} \\ \forall 1 \leq w \leq \mathrm{W} \\ \forall q \in \{0,1\} \end{matrix} \tag{4.22}$$

where Constraint (4.20) implies that a node v_n receives, at most, one column of $\mathscr{Z}_{\mathrm{W} \times \mathrm{U}}$ (i.e., one allocation pattern) when it acts as a sender. Also, Constraint (4.21) limits the bandwidth of v_n by the number requested by the transmitted flow. Constraint (4.22) deals with keeping interferences below the threshold level, which is a restatement of (4.2).

Finally, we propose the following objective function for the ILP model:

$$\begin{matrix} \max. \\ x_{ij}^{k,h}, t_n^k, c_{ij}, \\ \mathrm{H}_n, b_n, r_n^w, \cdots \end{matrix} \quad \alpha_\mathrm{R} \sum_{v_n \in \mathscr{V}} b_n + \alpha_\mathrm{A} \sum_{f^k \in \mathscr{F}} t_{s^k}^k - \alpha_\mathrm{N} \sum_{\substack{v_n \in \mathscr{V} \\ f^k \in \mathscr{F}}} \sum t_n^k \tag{4.23}$$

subject to: (4.3) – (4.22)

$$\text{with: } \alpha_\mathrm{R} \triangleq \frac{1}{\mathrm{W} \cdot |\mathscr{V}|}, \alpha_\mathrm{A} \triangleq \frac{1}{|\mathscr{F}_{\mathrm{WAIT}}|}, \alpha_\mathrm{N} \triangleq \frac{1}{|\mathscr{V}|}$$

Note this function is a weighted sum of the following optimization targets: i) maximize the RBs allocated to satisfy the bit rate demands, ii) maximize the number of offloaded flows, and iii) minimize the number of involved nodes.

As a final remark on the size complexity of the ILP model developed so far, we note that the model has row size (i.e., number of constraints) of $\mathscr{O}\left(|\mathscr{E}||\mathscr{F}| h_{\max} + |\mathscr{V}||\mathscr{E}|\mathrm{W}\right)$ and a column size (i.e., number of variables) of $\mathscr{O}\left(|\mathscr{E}||\mathscr{F}| h_{\max} + |\mathscr{V}||\mathscr{E}|\mathrm{W} + \mathrm{W}^2|\mathscr{V}|\right)$ assuming that $|\mathscr{E}| \approx |\mathscr{V}|$. Whereas ILP models are NP-hard to solve in general [73], this size complexity gives insight into the difficulty of solving this model to optimality and illuminates the reason behind the non-scalability of

JRW-D2D-NS.

4.2.2 Path-Based Formulation

In this sub-section, we propose an alternative formulation for the problem of joint routing and resource allocation. This formulation, which we call JRW-D2D-CG, is based on the Column-Generation approach for solving large-scale linear problems [74][75]. Instead of modeling a routing tree using **per-link variables**, one can map a tree to the union of simple paths where each path ends in a different destination.

In the following, we use the symbols p^k to denote a path from the set of all possible paths Π^k, over the topology $\mathbb{G}$, that carry the flow f^k to one of its destinations. We also introduce the symbol p to denote a path from the whole paths set $\Pi = \bigcup_{f^k \in \mathscr{F}} \Pi^k$. For brevity, we abuse the notation and: i) use the index n to refer to v_n, ii) use the ordered pair of indices ij to refer to e_{ij}, and iii) use the index k to refer to f^k.

To formulate the routing decision, we use a per-path 0-1 variable x_{p^k}, to show that the routing tree $\mathbb{T}^k$ includes the path p^k. We reintroduce the 0-1 variables t_n^k and c_{ij} defined earlier to make sure that the path selection agrees with the tree structure and keeps trees separated (i.e., the node-exclusivity). The resulting constraints are:

$$\sum_{p^k | \exists j | nj \in p^k} x_{p^k} - t_n^k \geq 0 \qquad \forall n \forall k \tag{4.24}$$

$$\mathsf{M} \cdot t_n^k - \sum_{p^k | \exists j | nj \in p^k} x_{p^k} \geq 0 \qquad \forall n \forall k \tag{4.25}$$

$$\sum_{p | ij \in p} x_p - c_{ij} \geq 0 \qquad \forall ij \tag{4.26}$$

$$\mathsf{M} \cdot c_{ij} - \sum_{p | ij \in p} x_p \geq 0 \qquad \forall ij \tag{4.27}$$

where M is a big constant which we set in each case accordingly. Then, we express the condition that these trees share no common node using the following constraint:

$$\sum_{k|n\notin\mathscr{D}^k} t_n^k + \sum_{k|n\in\mathscr{D}^k} t_n^{s^k} \leq 1 \qquad \forall n \forall k \tag{4.28}$$

Also, we guarantee the structure of routing trees using the following constraint:

$$\sum_{ij|j=n} c_{ij} \leq 1 \qquad \forall n \tag{4.29}$$

Moreover, as each selected routing tree implies the delivery of flow to all its destinations, we add:

$$\sum_{p^k|n=\mathsf{DEST}(p^k)} x_{p^k} - t_{s^k}^k \geq 0 \qquad \forall n \forall k \mid n \in \mathscr{D}^k \tag{4.30}$$

Regarding the assignment of half-duplex sets, we use the same 0-1 variables H_n's as in the link-based formulation, and apply the alternating link activation, using a similar constraint:

$$c_{ij} \leq \mathrm{H}_i + \mathrm{H}_j \leq 2 - c_{ij} \qquad \forall ij \tag{4.31}$$

However, for the RB allocation part, **we abandon the matrix formulation** used in the link-based formulation and propose a graph-based formulation since the latter leads to a more tractable problem than the matrix-based one. In this method, we represent an allocated RB by a selected link in a Resource-Block Allocation Graph (RBAG), as depicted in Figure 4.1. Using the RBAG, we encode the allocation of the RB w, by a 0-1 variable r_n^w that corresponds to the arc between the vertices representing the RB w and its successor $w+1$. Besides,

we use additional 0-1 variables y_n^w and z_n^w to encode the beginning and the end, respectively, of a RB allocation. These variables are attached to the virtual source and sink in the RBAG. We note that every contiguous allocation corre-

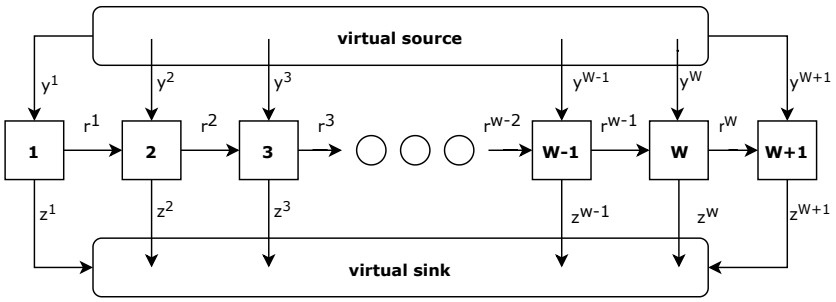


Figure 4.1 – RBAG: Resource Block Allocation Graph used to encode contiguous allocations.

sponds to a (continuous) path in the RBAG that starts from the virtual source and and ends in the virtual sink. Hence, we formulate the contiguity constraint on the RB allocation as:

$$\sum_{w=1}^{\mathrm{W}+1} y_n^w = \quad 1 \qquad \forall n \tag{4.32}$$

$$\sum_{w=1}^{\mathrm{W}+1} z_n^w = \quad 1 \qquad \forall n \tag{4.33}$$

$$y_n^w + r_n^{w-1} = r_n^w + z_n^w \quad \forall n \forall 1 \leq w \leq \mathrm{W}+1 \tag{4.34}$$

To represent the number of RBs allocated to a flow f^k, we use an integer variable b^k that must equal to the same bandwidth allocated to all nodes in the routing tree $\mathbb{T}^k$. Formally, we state this constraint as:

$$-\mathrm{M}\left(1-t_n^k\right) \leq b^k - \sum_{w=1}^{\mathrm{W}} r_n^w \leq \mathrm{M}\left(1-t_n^k\right) \quad \forall n \forall k \tag{4.35}$$

Besides, we limit the bandwidth of an accepted flow, by the respective bit rate,

using the following constraint:

$$b^k \leq \mathsf{RB}^k \cdot t^k_{s^k} \quad \forall k \tag{4.36}$$

We also reintroduce the 0-1 variable $r_n^{w,q}$ that determines if v_n transmits on the RB w during the opportunity of the half-duplex set $\mathcal{V}_{\mathrm{H}}^q$. However, we use the big-M formulation to avoid product terms as follows:

$$\sum_{w=1}^{\mathrm{W}} r_n^{w,q} \leq \mathsf{M}\big(1-q+(2q-1)\mathrm{H}_n\big) \qquad \begin{matrix}\forall q\in\{0,1\}\\ \forall n\end{matrix} \tag{4.37}$$

$$-\big(q+(1-2q)\mathrm{H}_n\big) \leq r_n^{w,q}-r_n^w \leq \big(q+(1-2q)\mathrm{H}_n\big) \qquad \begin{matrix}\forall q\in\{0,1\}\\ \forall n\\ \forall 1\leq w\leq \mathrm{W}\end{matrix} \tag{4.38}$$

Then, using the same big-M formulation, we rewrite the interference-management constraints as:

$$\Psi_\sigma r_i^{w,q} + \sum_{n\notin\{i,j\}} \Psi_t g_{nj} r_n^{w,q} \leq \frac{g_{ij}\Psi_t}{\Upsilon} + \mathsf{M}(2-r_i^{w,q}-c_{ij}) \qquad \begin{matrix}\forall q\in\{0,1\}\\ \forall ij\\ \forall 1\leq w\leq \mathrm{W}\end{matrix} \tag{4.39}$$

$$\begin{aligned} \max_{\substack{x_p,t_n^k,c_{ij},\\ \mathrm{H}_n,b^k,r_n^w,\cdots}} \quad & \mathsf{M}\sum_k b^k - \sum_{ij} c_{ij} \\ \text{subject to: } & (4.24)\text{ – }(4.39) \\ \text{with: } & \mathsf{M} > |\mathscr{E}| \geq \sum_{ij} c_{ij} \end{aligned} \tag{4.40}$$

Note that we must respect the condition on the constant M in (4.40) to make this single-objective optimization equivalent to a **lexicographical optimization** [76] in which we first maximize the total allocated RBs (i.e., max. $\sum_k b^k$)

and then, we minimize the involved relays (i.e., min. $\sum_{ij} c_{ij}$).

4.3 Proposal

In this section, we describe our proposal `JRW-D2D-CG` to solve the joint routing and resource block allocation problem using a variation of the delayed column-generation approach [74][75]. We note that in the extensive formulation of (4.40), it is hard to optimize over the entire set Π of possible paths (i.e., columns in the ILP model). As a matter of fact, the number of all possible paths $|\Pi|$ can be of $\mathcal{O}(|\mathcal{V}|!)$ in general. Instead, following the column-generation approach, we rely on the fact that most of these variables, namely the x_{p^k} 's, are zero in the final solution. So, we can find a method to generate, on-demand, only a subset of paths $\tilde{\Pi} \subseteq \Pi$ to consider in the main problem: this problem is known as the pricing (sub-)problem.

First, we restrict the domain of the Master Problem defined in (4.40) to an initial subset of paths $\tilde{\Pi}$, which contains only the shortest paths from each the source of each flow to one of its destination. Note that this Restricted Master Problem (RMP) is still an instance of ILP. To facilitate the pricing scheme, we relax the RMP to a Linear-Programming (LP) problem (i.e., ignoring the integrality restriction on its variables) and solve it using the well-known simplex method. Then, we propose a gradual build-up of the path subset $\tilde{\Pi}$. Using a pricing scheme for the paths in Π, we iteratively include only those that can improve the solution value of the relaxed RMP. However, it is possible to reformulate this pricing problem without explicitly iterating over the set Π.

To assess the improvements in the solution of the relaxed RMP, we need to calculate the Reduced Cost $\mathsf{RC}(\cdot)$ of the variable x_{p^k}, which is a function of the dual costs of constraints, in the current solution, and the coefficient of the

objective function. In our case, we have:

$$\begin{aligned}\mathsf{RC}(x_{p^k}) &= 0 - \sum_{n|\exists j|nj\in p^k}(\pi^{n,k}_{4.24} - \pi^{n,k}_{4.25}) - \sum_{ij|ij\in p^k}(\pi^{ij}_{4.26} - \pi^{ij}_{4.27}) - \pi^{\mathsf{DEST}(p^k),k}_{4.30} \\ &= \sum_{n|\exists j|nj\in p^k}(\pi^{n,k}_{4.25} - \pi^{n,k}_{4.24}) + \sum_{ij|ij\in p^k}(\pi^{ij}_{4.27} - \pi^{ij}_{4.26}) - \pi^{\mathsf{DEST}(p^k),k}_{4.30} \end{aligned} \tag{4.41}$$

where $\pi^{n,k}_{4.24}$, $\pi^{n,k}_{4.25}$, $\pi^{ij}_{4.26}$, $\pi^{ij}_{4.27}$, and $\pi^{n,k}_{4.30}$ are the dual costs of Constraints (4.24), (4.25), (4.26), (4.27), and (4.30) respectively. Since our problem is a maximization one, the attractive paths are those with $\mathsf{RC}(x_{p^k}) > 0$.

To populate the path subset $\tilde{\Pi}$, we propose an iterative pricing and populating scheme. After solving the relaxed RMP over the current subset $\tilde{\Pi}$, we look for a new path, for each flow f^k and each destination $d^k \in \mathscr{D}^k$, with the maximum favorable reduced cost. Once we find such paths, we add them to $\tilde{\Pi}$ and resolve the relaxed RMP again until we can no more find such paths. The pricing problem to find, a path p^k for the flow f^k and the destination $d^k \in \mathscr{D}^k$, can be stated formally as:

$$\max_{p^k\in\Pi^k_{n|\exists j|nj\in p^k}} \sum(\pi^{n,k}_{4.25} - \pi^{n,k}_{4.24}) + \sum_{ij|ij\in p^k}(\pi^{ij}_{4.27} - \pi^{ij}_{4.26}) \tag{4.42}$$

$$\text{such that:} \quad \mathsf{DEST}(p^k) = d^k$$

$$\sum_{n|\exists j|nj\in p^k}(\pi^{n,k}_{4.25} - \pi^{n,k}_{4.24}) + \sum_{ij|ij\in p^k}(\pi^{ij}_{4.27} - \pi^{ij}_{4.26}) > \pi^{d^k,k}_{4.30}$$

One can see that the pricing problem (4.42) is equivalent to finding the longest path in the weighted graph $\mathbb{G}$ from s^k to d^k, with node-weights $\pi^{n,k}_{4.25} - \pi^{n,k}_{4.24}$ for v_n, and link-weights $\pi^{ij}_{4.27} - \pi^{ij}_{4.26}$ for e_{ij}. This problem, under arbitrary weights, is known to be NP-hard [77].

To speed up the solution time, and to improve the scalability in case of dense topologies, we propose to solve the pricing problem over a restricted set

of paths. Every time we solve the pricing problem for a flow f^k and a destination d^k, we explore only a sub-graph of $\mathbb{G}$, which we call the decycled version relative to d^k. This decycled version is defined formally as:

$$\mathsf{DECYCLE}(\mathbb{G}, d^k) = \left\{ e_{ij} \in \mathbb{G} \mid \left| v_j - d^k \right| < \left| v_i - d^k \right| \right\} \tag{4.43}$$

It is straightforward to see that $\mathsf{DECYCLE}(\mathbb{G}, d^k)$ is the Directed Acyclic Graph (DAG) induced on $\mathbb{G}$ by a *topological ordering* of the nodes based on their distances from d^k. Thanks to the DAG property, we can solve the pricing problem (4.42), sub-optimally, over $\mathsf{DECYCLE}(\mathbb{G}, d^k)$, in $\mathrm{O}(|\mathcal{V}||\mathcal{E}|)$ worst-case time, using the Bellman-Ford algorithm [77] by negating the weights of vertices and links in the multi-version decycled graphs $\mathsf{DECYCLE}(\mathbb{G}, d^k)$ of $\mathbb{G}$. The last step in our method is to activate the integrality constraints in the RMP and to solve the ILP model over the final set $\tilde{\Pi}$ which gives, in general, a sub-optimal integer solution to the original master problem in (4.40). The pseudo-code in Algorithm 3 outlines the solution method of `JRW-D2D-CG`, described so far.

Algorithm 3 MILP resolution

Input: The ILP model of the MP in (4.40).

Output: A (sub-optimal) solution value for x_p, H_n, r_n^w, t_n^k, c_{ij}, b^k, $\cdots$.

$\tilde{\Pi} \leftarrow \emptyset$
for each $f^k \in \mathscr{F}$ and $d^k \in \mathscr{D}^k$ **do** ▷ initial paths
  $\tilde{\Pi} \leftarrow \tilde{\Pi} \cup \{\textsc{DijkstraShortestPath}(\mathbb{G}, s^k, d^k)\}$
end for
NewPaths $\leftarrow 1$
while NewPaths > 0 **do**
  NewPaths $\leftarrow 0$
  RMP $\leftarrow$ the MP restricted over $\tilde{\Pi}$
  LP $\leftarrow$ RMP with the integrality constraints removed
  Solve LP using the simplex algorithm
  Get the dual costs $\pi_{4.24}^{n,k}$, $\pi_{4.25}^{n,k}$, $\pi_{4.26}^{ij}$, $\pi_{4.27}^{ij}$, and $\pi_{4.30}^{n,k}$
  for each $f^k \in \mathscr{F}$ and $d^k \in \mathscr{D}^k$ **do**
    $\tilde{\mathbb{G}} \leftarrow \mathrm{DECYCLE}(\mathbb{G}, d^k)$ ▷ as defined in (4.43)
    Add the node-weights $\pi_{4.25}^{n,k} - \pi_{4.24}^{n,k}$ to $\tilde{\mathbb{G}}$
    Add the link-weights $\pi_{4.27}^{ij} - \pi_{4.26}^{ij}$ to $\tilde{\mathbb{G}}$
    $p_{\text{new}} \leftarrow \textsc{BellmanFordShortestPath}(-\tilde{\mathbb{G}}, s^k, d^k)$
    if p_{new} exists and $-\textsc{PathWeight}(p_{\text{new}}) > \pi_{4.30}^{d^k,k}$ **then**
      $\tilde{\Pi} \leftarrow \tilde{\Pi} \cup \{p_{\text{new}}\}$
      NewPaths $\leftarrow$ NewPaths $+ 1$
    end if
  end for
end while
RMP $\leftarrow$ the MP restricted over $\tilde{\Pi}$
Solve the ILP model RMP
return the solution value of x_p, H_n, r_n^w, t_n^k, c_{ij}, b^k, $\cdots$

4.4 Performance Evaluation

In this section, we give and discuss a performance evaluation of our proposed scheme `JRW-D2D-CG`. We base our evaluation on extensive network simulation in the NS-3 environment, which we augmented to support the LTE-D2D protocol stack as described in Section 3.4.

4.4.1 General Scenario Parameters

For our experiments, we use a scenario setup similar to the one described in Section 3.4 where we simulate a cellular network of a single non-sectorized cell with a radius of $R_{cell} = 1$ km which managed by one LTE-A eNB. Unless stated otherwise, we use the same scenario parameters in Table 3.1. Inside the cell, we distribute the nodes (i.e., the UEs) as a Poisson Point Process (PPP) with varying densities, λ_{UE} nodes per km^2, for values in the range $[10-100]$. For the evaluation, we set the confidence level to 95% to make assessments using the sample metrics.

4.4.2 Baselines for Comparison

To establish a baseline for comparison, we also simulate our original non-scalable scheme `JRW-D2D-NS` [40] for both simulated traffic scenarios. We also simulate another basic approach for routing and resource allocation, denoted by `DJK-RRB`, which is a simple routing scheme paired with random resource block allocation. Specifically, `DJK-RRB` finds the optimal routing trees (or paths) using the Dijkstra's shortest-path algorithm, and then it allocates the RBs randomly.

4.4.3 Collected Performance Metrics

To evaluate the performance of our proposed scheme JRW-D2D-CG as well as its competitor schemes, we use the metrics $\mathbb{S}$, $\mathbb{H}$, and $\mathbb{L}$ defined in Section 3.4.3 in addition to the following metrics:

- $\mathbb{L}$ is slightly modifed to consider the multicast case as follows:

$$\mathbb{L} = \mathrm{E}\left[\frac{\sum\limits_{f^k \in \mathscr{F}_{\mathrm{ADM}} \setminus \mathscr{F}_{\mathrm{INT}}} \frac{|\mathscr{D}^k|\mathtt{pkts}^k_{\mathrm{tx}} - \mathtt{pkts}^k_{\mathrm{rx}}}{|\mathscr{D}^k|\mathtt{pkts}^k_{\mathrm{tx}}}}{|\mathscr{F}_{\mathrm{ADM}}|}\right] \tag{4.44}$$

- $\mathbb{T}$ measures the average computation time required to solve one instance of the routing and resource allocation problem. This metric $\mathbb{T}$ is defined by:

$$\mathbb{T} = \mathrm{E}\left[\frac{\sum\limits_{\tau} \texttt{solution time of the frame } \tau}{\texttt{total frames in the simulation run}}\right] \tag{4.45}$$

Note that metric $\mathbb{S}$ accounts for the service admission ratio, which indicates the utility of the whole offloading system. On the other hand, metrics $\mathbb{H}$ and $\mathbb{L}$ are Quality-of-Service (QoS) related parameters. While metric $\mathbb{H}$ is related to end-to-end delays in packet transmissions, metric $\mathbb{L}$ is a measure of offloading service reliability. The solution-time metric $\mathbb{T}$ (i.e., convergence time) emphasizes the scalability of the offloading scheme gives some insights into its applicability in real-world settings.

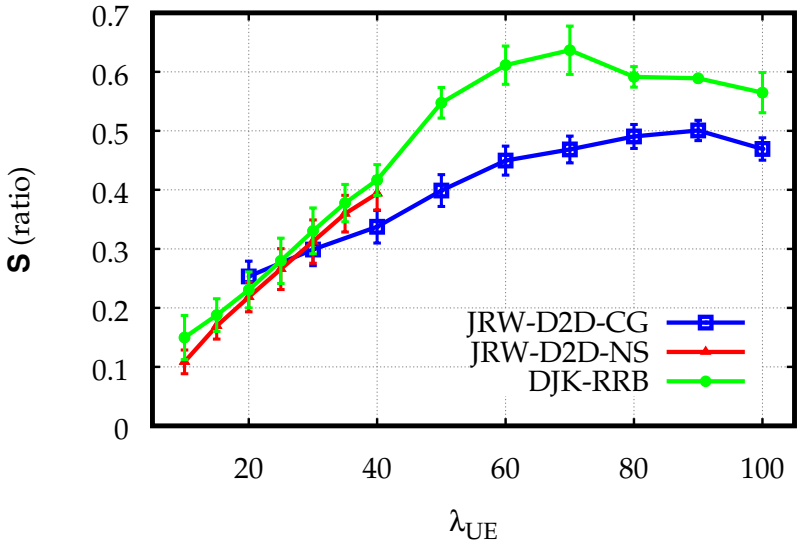
4.4.4 Unicast Applications Scenario

In this setup, we use a Poisson arrival process to generate unicast traffic flows. To reproduce different traffic load condition, we set the arrival rate of flows λ_{FL}

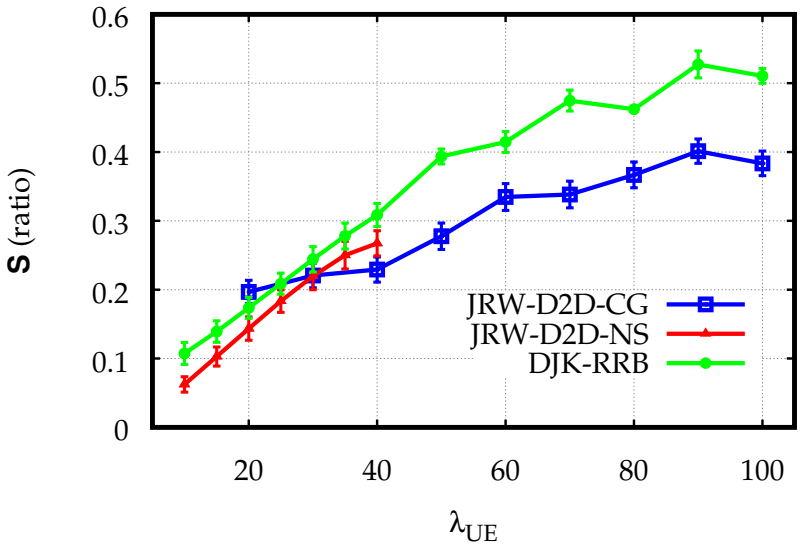
to values from $\{10, 20\}$ flows per second. Additionally, we suppose that the flows have different Constant Bit-Rate (CBR) selected randomly from the predefined classes shown in Table 3.1. As for the duration of flows, we assume that the values follow an exponential random variable with a mean duration of λ_{DUR} = 1 second. Finally, for the selection of communicating parties, we choose the sources and the destinations from a random uniform distribution.

4.4.5 Simulation Results in the Unicast Scenario

Figure 4.2 illustrates the degree of success of the offloading schemes at absorbing the unicast flows using the metric $\mathbb{S}$. Figure 4.2 (a) and Figure 4.2 (b) show the plots of $\mathbb{S}$ against the density of UEs distribution λ_{UE} under two different traffic load conditions. In both cases, we observe that, in general, the offloading capacity increases with the number of UEs. However, the fact that we cannot use a relay UE to route more than one flow at a time plays a role in how $\mathbb{S}$ evolves against UE. Particularly, employing more nodes to avoid busy nodes or interference zones reduces the offloading capacity. On the other hand, we can observe that the scheme `DJK-RRB` surpasses the other schemes in this regard. One can expect this performance since `DJK-RRB` is interference-agnostic and allocates fewer nodes. Nevertheless, as shown later, the decreased performance of `DJK-RRB` for other metrics outweighs this remarkable advantage. For the schemes `JRW-D2D-CG` and `JRW-D2D-NS`, we remark that the non-scalable `JRW-D2D-NS` slightly outperforms `JRW-D2D-CG` for low-density UE deployments (i.e., $\lambda_{UE} \leq 30$). This small advantage is due to the sub-optimality of `JRW-D2D-CG`, which speeds up the solution process by considering only the paths over the multi-version decycled graphs of the topology, as explained earlier. One should also remark that we have lower acceptance ratios $\mathbb{S}$ in the traffic condition λ_{FL} = 20 than those in λ_{FL} = 10. This means that we have already reached



(a) $\lambda_{FL} = 10$

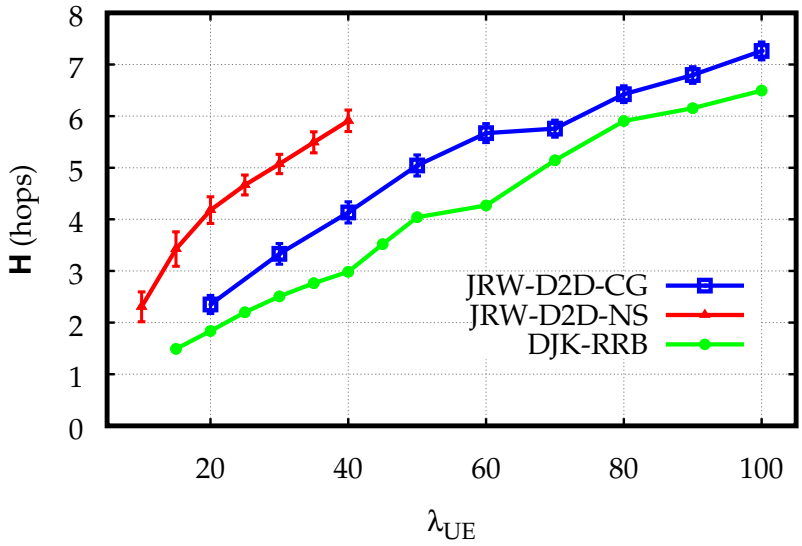


(b) $\lambda_{FL} = 20$

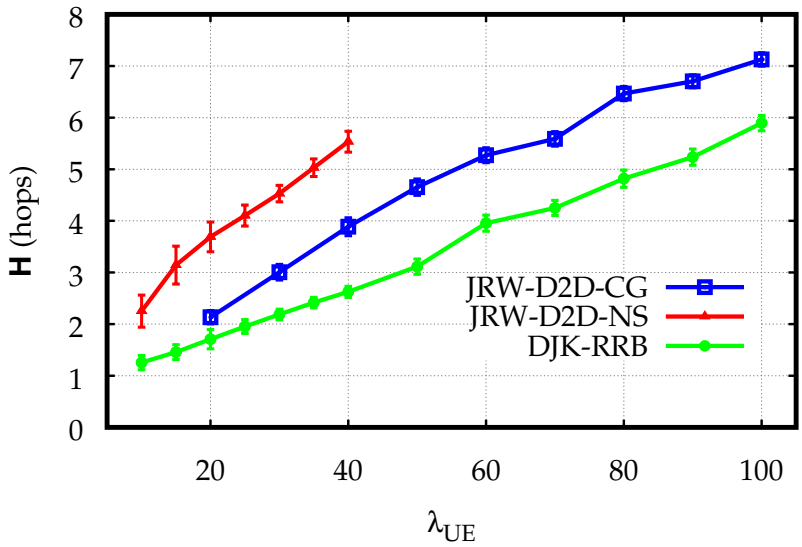
Figure 4.2 – The offloading ratio metric $\mathbb{S}$ versus the nodes density λ_{UE}.

the offloading capacity of the system.

Figure 4.3 illustrates the plots of the metric $\mathbb{H}$ against λ_{UE} for two differ-



(a) $\lambda_{FL} = 10$



(b) $\lambda_{FL} = 20$

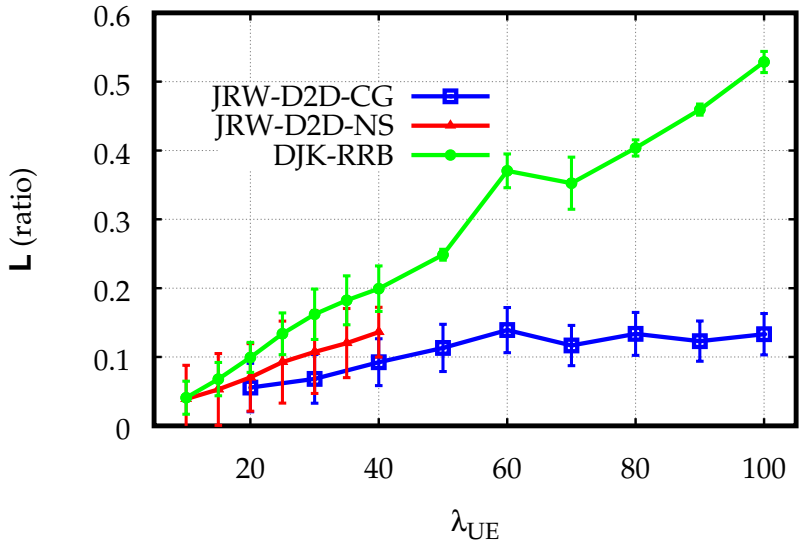
Figure 4.3 – The hop metric $\mathbb{H}$ versus the nodes density λ_{UE}.

ent traffic condition. As said before, smaller values for $\mathbb{H}$ (i.e., shorter routes) mean shorter end-to-end delays and less involved UEs. In the two traffic con-

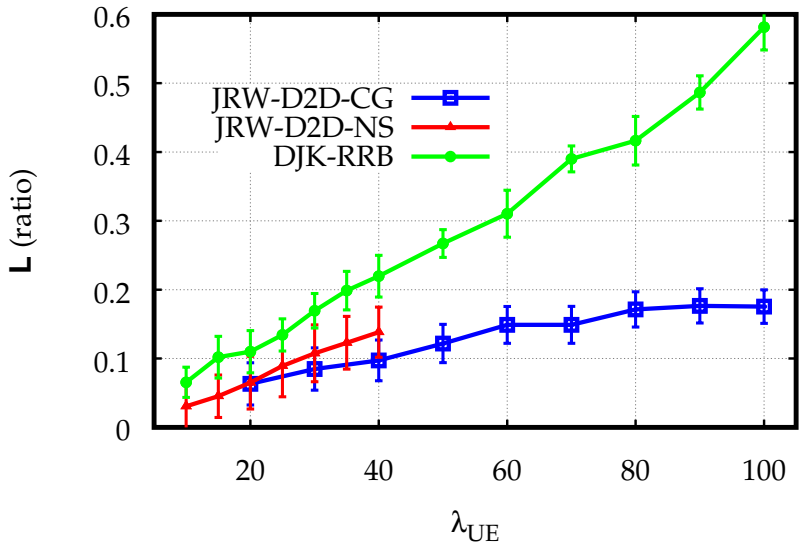
ditions in Figure 4.3 (a) and Figure 4.3 (b), we notice that DJK-RRB generates, as expected, the shortest possible routes. However, we note that JRW-D2D-CG is, at most, one-hop off DJK-RRB. Moreover, JRW-D2D-CG clearly outperforms the non-scalable variant JRW-D2D-NS in the range $\lambda_{UE} \leq 40$ beyond which JRW-D2D-NS fails to give solutions in a reasonable time. Likewise JRW-D2D-CG produces shorter routes than JRW-D2D-NS because, as said before, JRW-D2D-CG only considers the multi-version decycled graph.

In Figure 4.4 (a) and Figure 4.4 (b), the plots of $\mathbb{L}$ against λ_{UE} show that JRW-D2D-CG offers more reliable offloading service than DJK-RRB. This advantage is remarkably significant in high-density UE deployments in which DJK-RRB causes a high level of interference, whereas JRW-D2D-CG can keep the packet loss rate under 0.2. Moreover, JRW-D2D-CG gives nearly the same performance as the non-scalable solved-for-optimality JRW-D2D-NS in low-density deployments.

With regards to scalability, Figure 4.5 shows the plots of the metric $\mathbb{T}$ (i.e., computation time) versus λ_{UE}. The plots demonstrate that the non-scalable JRW-D2D-NS takes a significantly longer time than JRW-D2D-CG to solve one instance of the optimization problem. JRW-D2D-NS begins to struggle with topologies with $\lambda_{UE} > 20$ as it takes on average around 100 seconds to solve instances of $\lambda_{UE} = 40$. In contrast, in less than 100 seconds, JRW-D2D-CG can easily solve instances of λ_{UE} values up to 100. Hence, JRW-D2D-CG is more scalable than JRW-D2D-NS. This result demonstrates the effectiveness of solving the pricing problem over the multi-version decycled graph using the fast Bellman-Ford algorithm.



(a) $\lambda_{FL} = 10$



(b) $\lambda_{FL} = 20$

Figure 4.4 – The packet loss metric $\mathbb{L}$ versus the nodes density λ_{UE}.

4.4.6 Multicast Applications Scenario

In addition to the general scenario assumptions, we keep the same simulation parameters as the unicast scenario regarding the Poissonian flow arrival,

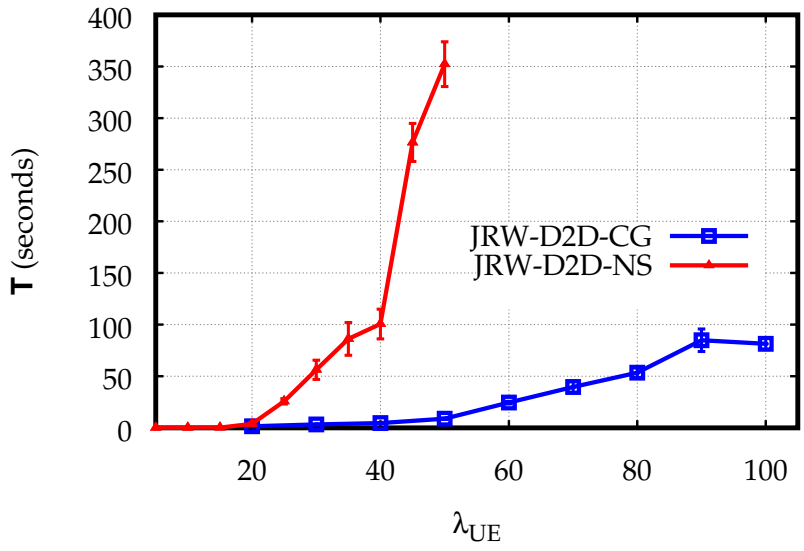
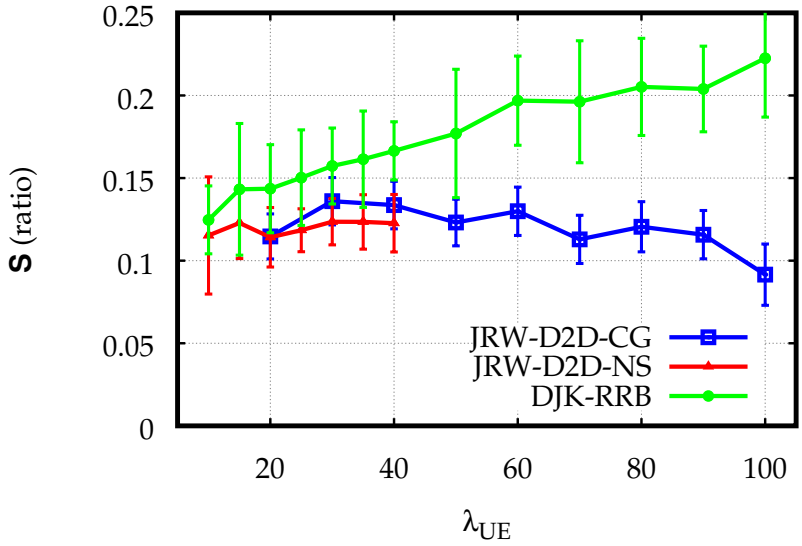


Figure 4.5 – The computation time metric $\mathbb{T}$ versus the nodes density λ_{UE}.

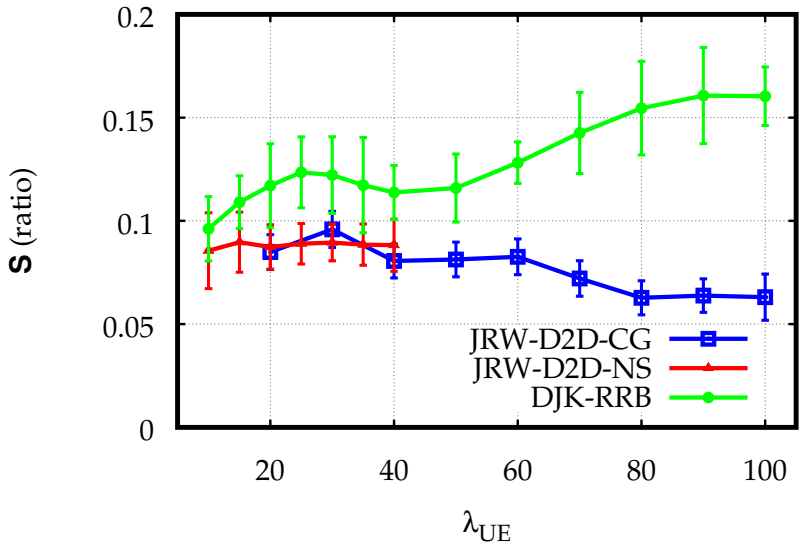
the flow duration, and the bit rates. However, for the selection of the sources and destinations, we use a fixed subscription-rate to model the probability of a node to subscribe as a receiver. In this scheme, we select the sources randomly following a uniform distribution. Moreover, as for destinations, we assume a node-flow interest probability $\rho = 0.1$, and we decide whether every other node is a receiver using Bernoulli trials with a success probability of ρ.

4.4.7 Simulation Results in the Multicast Scenario

Figure 4.6 shows the plots of the metric $\mathbb{S}$ against λ_{UE} under the multicast traffic condition. Similar to the unicast case, we observe that `DJK-RRB` performs better than both `JRW-D2D-CG` and `JRW-D2D-NS` for the same reasons explained earlier. Moreover, both `JRW-D2D-CG` and `JRW-D2D-NS` give essentially similar performances for $\lambda_{UE} \leq 40$. We also observe that $\mathbb{S}$ does not significantly change when λ_{UE} increases as it is the case for the unicast scenario. This can be re-



(a) $\lambda_{FL} = 10$



(b) $\lambda_{FL} = 20$

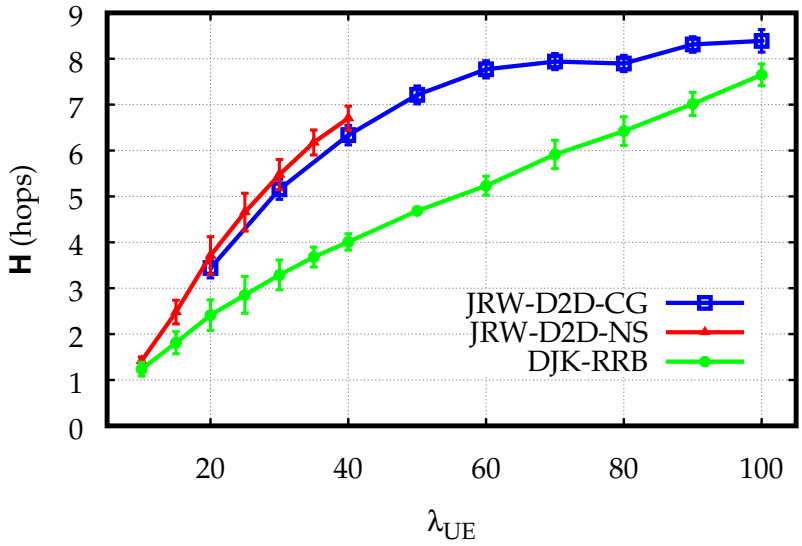
Figure 4.6 – The offloading ratio metric $\mathbb{S}$ versus the nodes density λ_{UE}.

garded as an artifact of how we select the destinations of the simulated flows. Namely, the multicast group size is proportional to the number of UEs with an

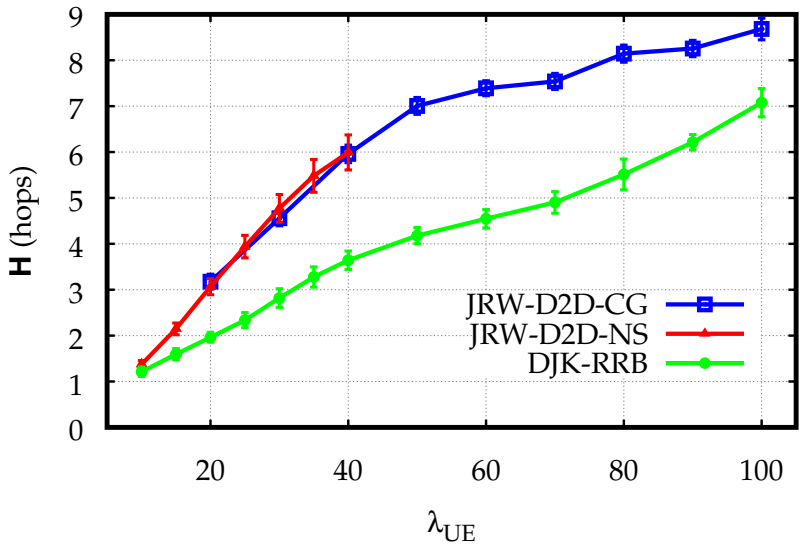
average of ρN due to using Bernoulli trials to assign the destinations. Consequently, the average number of simultaneous flows is about $\frac{1}{\rho}$. In other words, the fixed subscription-rate model, used in the traffic model, puts a limit on the offloading capacity regardless of the number of UEs in the network. Nonetheless, Figure 4.6 (a) and Figure 4.6 (b) show that JRW-D2D-CG is capable to serve at least 10% and 5% of the flows under the traffic load conditions $\lambda_{FL} = 10$ and $\lambda_{FL} = 20$ respectively.

Figure 4.7 illustrates the behavior of the metric $\mathbb{H}$ against λ_{UE} in the multicast situation. Similar to the unicast setup, as Figure 4.7 (a) and Figure 4.7 (b) demonstrate, the average route length increases with the topology size. We also observe that DJK-RRB outdoes both JRW-D2D-CG and JRW-D2D-NS as implied by its definition, as explained previously. As for performance gaps, we note that JRW-D2D-CG gives average routes lengths, which are, at most, two hops longer than those of DJK-RRB. Moreover, as before, JRW-D2D-CG performs as good as JRW-D2D-NS for $\lambda_{UE} \leq 40$.

Figure 4.8 describes the plot of metric $\mathbb{L}$ versus λ_{UE} in the multicast setup. We can make similar statements like those in the unicast setup. We note that DJK-RRB, being agnostic to interference, offers less reliable end-to-end transmission than the interference-aware schemes JRW-D2D-CG and JRW-D2D-NS. Moreover, we note that, under the two traffic conditions which Figure 4.8 (a) and Figure 4.8 (b) represent, JRW-D2D-CG does much better than JRW-D2D-NS in the range for $\lambda_{UE} \leq 40$ demonstrating that routing over the multi-version decycled graph produces better multicast trees with regards to interference. In this regards, JRW-D2D-CG keeps the packet loss rate under 0.3. We also note that the high loss rate, relative to the unicast setup, is due to the higher number of interfering links in routing trees compared to the unicast case where routes are simple paths.



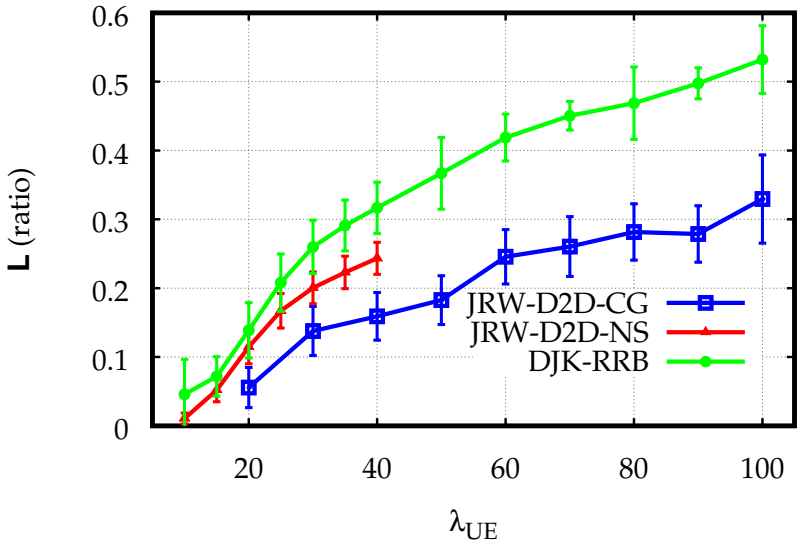
(a) $\lambda_{FL} = 10$



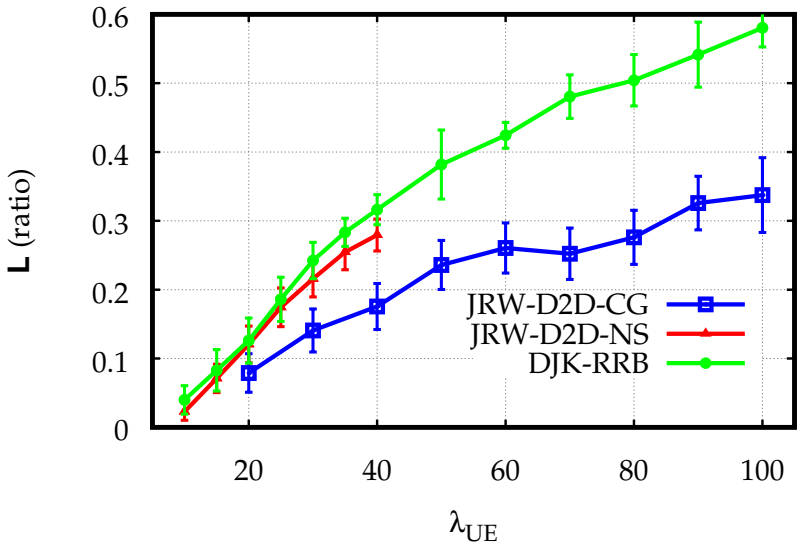
(b) $\lambda_{FL} = 20$

Figure 4.7 – The hop metric $\mathbb{H}$ versus the nodes density λ_{UE}.

To sum up, we can conclude from the above extensive network simulation results that our novel scheme `JRW-D2D-CG` leads to the best system per-



(a) $\lambda_{FL} = 10$



(b) $\lambda_{FL} = 20$

Figure 4.8 – The packet loss metric $\mathbb{L}$ versus the nodes density λ_{UE}.

formance in terms of reliability and service (admission) ratio, and more importantly, the scalability to systems with high-density UE deployments. Sim-

ulations also demonstrate the usefulness of the techniques employed in JRW-D2D-CG to work around the formulation elements of JRW-D2D-NS that limit the scalability.

4.5 Conclusion

In this chapter, we discussed designing offloading schemes based on LTE-D2D to offload traffic, both unicast and multicast, within an LTE-A omnidirectional macro-cell. We addressed the main problem of joint routing and OFDMA resource allocation that underlies such schemes. We presented two ILP formulations for the problem that consider interferences between D2D links and low-level details of LTE-D2D technology: namely the contiguous resource block and the half-duplex operation of D2D links. The initial link-based formulation, named JRW-D2D-NS, does not scale well for large-scale D2D topologies because it solves the ILP model to optimality. Next, we presented a new path-based formulation called JRW-D2D-CG, where we used a method based on column-generation to solve the resulted ILP model in a sub-optimal way for reasons of speed. The novel formulation also included a novel and more tractable graph-based approach to the resource block allocation. Network simulations using NS-3 demonstrated that our novel proposal is more scalable than the original one. Furthermore, in general, both proposals have similar performances in terms of the reliability and latency of the offered service.

Offloading schemes, such as the proposals presented so far, do not consider the energy issue in relaying through UEs, which are usually battery-limited terminals. A practical Offloading scheme, based on the collaboration of UEs, should consider this battery issue to avoid interruptions in the offloading service. Moreover, Lowering the impact of data forwarding on battery life also helps motivate

users to participate in the whole system.

Chapter 5

D2D-Based Cellular Traffic Offloading: An Energy-Aware Scalable Heuristic Scheme

Contents

Data offloading based on LTE-D2D can support congestion-prone cellular networks in the face of traffic growth. In this chapter, we tackle the design of offloading systems that are aware of the energy limitation of the complementary LTE-D2D network. The general idea is to design a scheme that exploits a network of D2D-connected User-Equipments (UEs) to carry intracellular traffic relieving the eNB from the data plane's burdens. Nevertheless, the eNB man-

ages this offloading process by centrally executing the routing decisions and the frequency resource allocation. We also assume that the eNB is also aware of the energy budget of every participating UE. In this chapter, we present two approaches for the eNB to optimize this centralized operation, given its capability to have a global view of the system. The former is an optimal approach based on Integer-Linear-Programming (ILP) that does not scale well with high-density topologies. The second approach, named HERRA, is a novel heuristic scheme that uses a parametric three-stage method that includes possible variations on the strategies employed in its stages. Performance evaluation, using network simulations in NS-3, shows that our new proposal, HERRA, outperforms the original one in the matter of convergence time. As a result of massive speedups, up to six orders of magnitude, HERRA scales very well in denser topologies at the price of having some performance gaps, particularly in terms of packet loss.

5.1 Introduction

In the previous two chapters, we considered the problem of routing and RB allocation in the design of an offloading system for UE-to-UEs traffic based on a multihop cooperative relaying between LTE-D2D UEs. We considered both unicast and multicast, and we also tackled the issue of scalability. In this chapter, we address the problem of routing and frequency resource allocation considering the energy and the scalability issues in the design of centralized algorithms in high-density deployments of UEs.

As assumed before, under high traffic-load condition, one eNB tries to command the UEs to deliver some of the intracellular flows from the source UEs to the destination UEs using multiple hops of D2D links (i.e., SLs). The eNB

takes charge of finding routes and allocating frequency resources for the concurrent D2D transmissions over the SL. To increase the utility of the offloading system, the eNB also considers the limited battery of the UE relays using an energy budget reserved for the offloading operation. Moreover, the eNB must run an algorithm that considers the LTE-D2D specific constraints. These include the half-duplex nature of the D2D interface and the fact that the granted Resource Blocks (RBs) for UEs, to use in SL, must be contiguous in the frequency domain.

Concerning the reviewed related work, we note that [56] considers only UE-to-BS traffic where high-battery UEs help low-battery analogs to relay their traffic to the BS. Moreover, our work focuses on offloading UE-UEs traffic (multicast or unicast) to relieve the base-station. Similarly, [57] also tackles the BS-to-UEs multicast video traffic where UEs employ a distributed multipath routing and caching technique. First, in this chapter, we put forward a scheme for the routing and the RB allocation for an energy-efficient offloading mechanism within the LTE networks. The scheme can handle both multicast and unicast flow-oriented applications uniformly. Precisely, we design within an LTE-D2D based offloading system that employs a sub-network of (LTE-D2D-enabled) UEs to route flows that begin and end in the same macro-cell. Our presented design, while being energy-budget aware like [57], focuses on central algorithms and flow-centric applications where the on-demand cluster formation and caching, in [57], cannot be used. Likewise, the protocol in [58] cannot be adapted for LTE-D2D to serve our purpose, since the traffic model in WSN is multiple-source, one-sink, and the clustering technique is ineffective in our case. The work in [59] is the closest to this chapter despite its emphasis on unicasting UE-UE traffic. However, this work makes generic assumptions about wireless technology, and the medium is abstracted as whole channels, not in

terms of RBs. Although, the work also incorporates an analytical power consumption model for each D2D link, however, it does not consider the energy-budget limitation. The first proposal of this chapter, in contrast, considers UE-to-UEs traffic where unicast traffic can be handled as a particular case. Furthermore, we recognize the specific constraints of LTE-D2D, and we employ an empirical power consumption model to address the efficient use of the energy budgets assigned for the collaborative offloading. In the second proposal of this chapter, we present another energy-aware offloading scheme, called HERRA, to solve the routing and RB allocation that meets the prior requirements. However, unlike the first one, the new scheme scales very well to larger topologies.

The most notable contributions of this chapter are summarized as follows:

- We introduce an initial energy-aware offloading scheme, named JRRA-EA, based on an Integer Linear Programming (ILP) model. JRRA-EA jointly optimizes the routing and RB allocation in the system. We also give highlights on its complexity and non-scalability.

- We present a novel parametric heuristic-based energy-aware offloading scheme, named HERRA, to solve the routing and the RB allocation problems. In addition to being parameter-dependent, the given HERRA scheme is also paired with different strategies.

- Using extensive simulations in the network simulator NS-3, which we augmented to support LTE-D2D, we evaluate the performance HERRA with its different variations and compare them to the first proposal JRRA-EA.

Simulation results demonstrate that our novel proposal HERRA converges faster than the original non-scalable JRRA-EE achieving massive speedups. Owing to this, the proposal HERRA scales very well to high-density deployments of D2D nodes. Furthermore, HERRA is more flexible to further improvements due

to having variations and being parameter-dependent, which leaves more room for enhancement towards additional performance targets. However, HERRA's scalability comes at the price of losing some optimality, relative to JRRA-EE, in terms of service reliability.

The remainder of this chapter is organized as follows. In Section 5.2, we provide an in-depth description of our network model, including the ILP-based scheme, to solve the problem mentioned above. Next, Section 5.3 presents our new heuristic proposal to solve the same problem in a more tractable way. In Section 5.4, we discuss the network simulation results and give a comparative evaluation of our proposal relative to the exact resolution scheme JRRA-EE. Lastly, Section 5.5 ends the chapter by summarizing the principal results.

5.2 Network Model

This section introduces the network model for which we conceive our energy-aware offloading scheme. We consider a cellular region covered by one LTE-A macro-cell (eNB), which serves N D2D-ready user-terminals (UEs). Considering these densely collocated UEs, we assume a scenario where they generate a high load of UE-to-UE data traffic between them. In the general case, we consider one-to-many (i.e., multicasting) flows. Each flow has a source UE and a group of destination UEs. To avoid congestion, we assume the UEs are ready to offload this type of traffic while being under the control of the eNB. To this end, each UE reserves an energy budget to participate in the offloading process. Besides, the eNB has a global view of the topology composed of the viable D2D links (SLs) between the UEs and perfect knowledge of the energy budget of each UE. We also assume that the D2D topology is stable during the high-load period. In other words, the UEs do not move substantially. Such a topology

of links can offer an offloading service using an always-on connected relaying mechanism. Such an arrangement greatly facilitates the design of the overall system. Because no sophisticated buffer management is needed as in opportunistic store-and-forward relays and hence it reduces the amount of signaling needed for the eNB to control the whole operation.

In order for such an offloading scheme to be useful, the eNB must solve the following problems: i) which flows to admit into the offloading system? ii) what are the optimal routes from sources to destinations?, iii) given the dedicated number of OFDMA RBs to the offloading and the half-duplex nature of the LTE-D2D interface, how can the D2D links be scheduled simultaneously without causing harmful interference to each other? and iv) how to minimize the service disruption due to the energetic death of relays? Also, the eNB must consider the energy-budget issue of the system to increase its utility, so it performs longer.

To formulate the problem, we continue to use the same formalism of Section 3.2 and Section 4.2. Same as before, we model the topology using a directed graph $\mathbb{G}=(\mathcal{V},\mathcal{E})$. The set of nodes $\mathcal{V}$ represents the N UEs, and each edge e_{ij} from the set $\mathcal{E}$ represent a viable communication link between the respective nodes (i.e., UEs) v_i and v_j. Note that the eNB discovers a viable link from v_i to v_j if the respective Signal-to-Noise Ratio (SNR), γ_{ij}, is higher than a predefined threshold, γ_{TOPO}.

For the link activation, we use the same alternating link activation described in Chapter 3 and Chapter 4. Accordingly, the eNB classifies the nodes in the topology into: i) the active half-duplex sets $\mathcal{V}_{\text{H}}^{p}$ for $p \in \{0,1\}$ where $\mathcal{V}_{\text{H}}^{0}$, $\mathcal{V}_{\text{H}}^{1}$ are the sets of nodes to transmit in the even and odd frame-sets respectively, ii) the idle nodes $\mathcal{V}_{\text{D}}$, and iii) the $\mathcal{V}_{\text{X}}$ the set of departed nodes (or dead nodes) which have already exhausted their energy budget for the relaying process.

In line with the formulations of Chapter 3 and Chapter 4, the eNB can allocate the same RB to different nodes if it can keep the mutual interference below a harmful level. To do this, we assume the same per-RB treatment of Section 3.2, Section 3.2. Similarly, we assume the same fixed emission power density Ψ_t [mW/RB], for all the nodes, and the same flat block-fading channel model.

Offloading the flow f^k means that the eNB must find a route $\mathbb{T}^k$, which is generally a tree, over the idle nodes from the source to the destination(s). Besides, the eNB must also decide about the half-duplex set assignments of the reserved nodes. Moreover, it must then continuously allocate enough RBs to these nodes in each frame. Note that these decisions should be optimal in some sense, as defined later. Thus, in each frame, the eNB classifies all the flows in the system, besides the finished flows, into i) the set $\mathscr{F}_\mathrm{S}$ of ongoing (scheduled) flows, and ii) the set $\mathscr{F}_\mathrm{W}$ of flows waiting to be offloaded. The eNB can also filter $\mathscr{F}_\mathrm{W}$ down to a set $\mathscr{F}_\mathrm{C} \subseteq \mathscr{F}_\mathrm{W}$ of candidate flows to be considered for the offloading the next frame.

Empirical Power Consumption Model for UEs

To estimate the energy consumption, similar to [78], we make use of the empirical model of UE defined in [79] which calculates the energy consumption due to the D2D direct communication at both endpoints. For a UE transmitting data, the power consumption $\mathrm{P}_{\mathrm{tx}}^{\mathrm{D2D}}$ (mW), using the values of the model parameters in Table 5.1, is estimated as:

$$\begin{aligned} \mathrm{P}_{\mathrm{tx}}^{\mathrm{D2D}} &= \mathrm{P}_{\mathrm{tx}}^{\mathrm{const}} + \mathrm{P}_{\mathrm{tx}}^{\mathrm{RF}}(\mathrm{S}_{\mathrm{tx}}) \\ \mathrm{P}_{\mathrm{tx}}^{\mathrm{RF}}(\mathrm{S}_{\mathrm{tx}}) &= \begin{cases} b_1^{\mathrm{tx}} \cdot \mathrm{S}_{\mathrm{tx}} + a_1^{\mathrm{tx}} & \text{if } \mathrm{S}_{\mathrm{tx}} \leq s_1^{\mathrm{tx}} \\ b_2^{\mathrm{tx}} \cdot \mathrm{S}_{\mathrm{tx}} + a_2^{\mathrm{tx}} & \text{if } s_1^{\mathrm{tx}} < \mathrm{S}_{\mathrm{tx}} \leq s_2^{\mathrm{tx}} \end{cases} \end{aligned} \tag{5.1}$$

Table 5.1 – Parameters of the UE Power Consumption Model

Parameter	Value
P_{tx}^{const}	883.52 mW
P_{rx}^{const}	878.1 mW
s_1^{tx}	0.2 dBm
s_2^{tx}	11.4 dBm
s_1^{rx}	52.5 dBm
a_1^{tx}	23.6 mW
a_2^{tx}	45.4 mW
a_1^{rx}	24.8 mW
a_2^{rx}	7.86 mW
a^{R}	8.16 mW
b^{R}	0.97 mW/Mbps
b_1^{tx}	0.78 mW/dBm
b_2^{tx}	17 mW/dBm
b_1^{rx}	0.04 mW/dBm
b_2^{rx}	0.11 mW/dBm

where P_{tx}^{const} is the power consumption of the baseband circuit when the transmitter is active, and P_{tx}^{RF} is the power consumption of the whole RF block as a function of the power emitted S_{tx} from the antenna in dBm.

For a UE actively receiving data, and using the model parameters values in Table 5.1, the power consumption P_{rx}^{D2D} (mW) is estimated as:

$$\begin{aligned} P_{rx}^{D2D} &= P_{rx}^{const} + P_{rx}^{RF}(S_{rx}) + P_{rx}^{BB}(R) \\ P_{rx}^{RF}(S_{rx}) &= \begin{cases} -b_1^{rx} \cdot S_{rx} + a_1^{rx} & \text{if } S_{rx} \leq -s_1^{rx} \\ -b_2^{rx} \cdot S_{rx} + a_2^{rx} & \text{if } S_{rx} > -s_1^{rx} \end{cases} \\ P_{rx}^{BB}(R) &= b^{R} \cdot R + a^{R} \end{aligned} \tag{5.2}$$

where P_{rx}^{const} is the power consumption of the baseband circuit when the receiver is active, and P_{rx}^{RF} is the power consumption of the whole RF block as a function of the power received S_{rx} from the antenna in dBm. As for the addi-

tional term P_{rx}^{BB}, it estimates the power consumption in the baseband circuitry of the device, which is data-rate dependent.

Based on this empirical model, the power consumption of the node v_n that transmits the data of the flow f^k, $\Pi_{tx,n}^k$, can be estimated as:

$$\begin{aligned} \Pi_{tx,n}^k &= P_{tx}^{const} + P_{tx}^{RF}\left(S_{tx,n}^k\right) \qquad [mW] \\ S_{tx,n}^k &= dBm\left(\Psi_{tx,n} \cdot RB^k\right) \end{aligned} \tag{5.3}$$

Similarly, at the receiver side v_j of an active link e_{ij} with a link gain g_{ij}, which receives the data of the flow f^k, the power consumption of v_j, $\Pi_{rx,ij}^k$, is estimated by:

$$\begin{aligned} \Pi_{rx,ij}^k &= P_{rx}^{const} + P_{rx}^{RF}\left(S_{rx,ij}^k\right) + P_{rx}^{BB}\left(R^k\right) \qquad [mW] \\ S_{rx,ij}^k &= dBm\left(g_{ij} \cdot \Psi_{tx,i} \cdot RB^k\right) \end{aligned} \tag{5.4}$$

Thanks to the model above, the eNB can track the time evolution of the residual energy budget for a node v_n, $E_n(\tau)$, as follows:

$$E_n(\tau) = E_n(\tau - 1) - P^{D2D} \cdot T_{SL} \tag{5.5}$$

$$P^{D2D} = \begin{cases} P_{tx}^{D2D} \text{ if } v_n \text{ was transmitting in the frame } \tau - 1 \\ P_{rx}^{D2D} \text{ if } v_n \text{ was receiving in the frame } \tau - 1 \\ 0 \qquad \text{if } v_n \text{ was idle in the frame } \tau - 1 \end{cases}$$

where $E_n(\tau)$ is the residual energy budget of the node v_n at the beginning of the frame τ.

ILP-Based Problem Formulation

In this section, we present an initial offloading scheme to optimize the offloading decision in the eNB, as described above. This scheme, named JRRA-EE, employed an ILP formulation to jointly optimize the routing and the RB allocation in addition to being energy budget aware. We propose the joint treatment of routing and the RB allocation as a cross-layer optimization. We aim to obtain more optimal results when the routing also considers the induced interferences and the energy consumed in the communication endpoints. Moreover, we propose a batch mode to optimize the decision considering all the flows in the waiting queue instead of flow-by-flow decisions.

In the ILP model, we represent the routing, half-duplex assignment, and RB allocation by essential 0-1 decision variables $x_{ij}^{h,k}$, H_n, and r_n^w, respectively. The variables $x_{ij}^{h,k}$, for h=0, 1, $\cdots$, $h_{\max}$, give the constructed route $\mathbb{T}^k$ (generally, a tree) for the flow f^k over the (idle) nodes in the topology. A link e_{ij} is selected to be a part of $\mathbb{T}^k$ at the (tree) level h only when $x_{ij}^{h,k}$=1. An example of such route construction is shown in Figure 5.1.

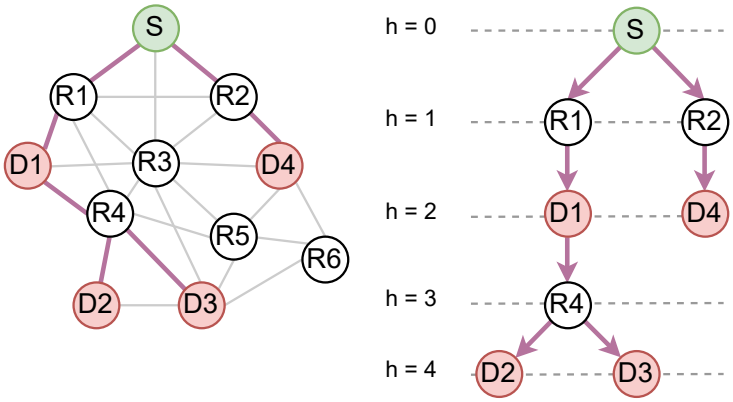


Figure 5.1 – Example of finding a route (tree) over a topology.

However, to ensure that the link selections are compatible with the mathe-

matical structure of a tree (or a simple path), we must impose more constraints on the variables $x_{ij}^{h,k}$. Let $\mathcal{O}(v_n)$ and $\mathcal{T}(v_n)$ be the sets of outgoing and incoming edges (links), respectively, at $v_n \in \mathbb{G}$. Then, we formulate the mentioned constraints as follows. We impose that each node has at most one parent (predecessor) node in a route by:

$$\sum_{\substack{e_{ij}\in\mathcal{T}(v_n)\\ 0\le h\le h_{\max}, f^k\in\mathcal{F}}}\sum\sum x_{ij}^{h,k} \le 1 \qquad \forall v_n\in\mathcal{V} \tag{5.6}$$

The last constraint also ensures that an edge cannot participate in more than one route (i.e., flow). Also, we express the fact that only edges from the source node are allowed at the root of the tree (i.e., at h=0) by the following constraint:

$$x_{ij}^{h,k} \le \delta_0^h \cdot \delta_{v_i}^{s^k} + (1-\delta_0^h)(1-\delta_{v_i}^{s^k}-\delta_{v_j}^{s^k}) \quad \begin{matrix}\forall e_{ij}\in\mathcal{E}\\ \forall 0\le h\le h_{\max}\\ \forall f^k\in\mathcal{F}\end{matrix} \tag{5.7}$$

where we used the Kronecker delta function δ_y^x, which equals to 1 only when x=y, to have a compact notation. Also, we enforce the continuity of the route by stipulating that an outgoing edge is allowed at the level h only when a predecessor edge is selected at the level $h-1$. We formally express this constraint using the following inequality:

$$x_{nm}^{h,k} \le \sum_{e_{ij}\in\mathcal{T}(v_n)} x_{ij}^{h-1,k} \quad \begin{matrix}\forall e_{nm}\in\mathcal{E}\\ \forall 1\le h\le h_{\max}\\ \forall f^k\in\mathcal{F}\end{matrix} \tag{5.8}$$

It is straightforward to see that each directed route formed according to the previous constraints is a non-circular graph. Besides, we must add the requirement that routes end only at destination nodes. In terms of the tree structure, we require that only a destination can be a leaf in a tree. Using the set indicator

function $\mathbb{1}_x^{\mathrm{Y}}$, which equals to 1 only when $x \in \mathrm{Y}$, we state this constraint formally as:

$$\sum_{\substack{e_{ij} \in \mathcal{T}(v_n) \\ 0 \le h \le h_{\max}}} \sum x_{ij}^{h,k} - \sum_{\substack{e_{ij} \in \mathcal{O}(v_n) \\ 0 \le h \le h_{\max}}} \sum x_{ij}^{h,k} \le \mathbb{1}_{v_n}^{\mathscr{D}^k} \qquad \begin{matrix} \forall v_n \in \mathcal{V} \\ \forall f^k \in \mathscr{F} \end{matrix} \tag{5.9}$$

To proceed with the formulation, we define an auxiliary 0-1 variable t_n^k which indicates whether the node v_n acts a transmitter in the route $\mathbb{T}^k$ (for the flow f^k). The following constraints fix this variable in terms of the variables $x_{ij}^{h,k}$:

$$t_n^k \ge \sum_{0 \le h \le h_{\max}} x_{nm}^{h,k} \qquad \begin{matrix} \forall e_{nm} \in \mathcal{E} \\ \forall f^k \in \mathscr{F} \end{matrix} \tag{5.10}$$

$$t_n^k \le \sum_{\substack{e_{ij} \in \mathcal{O}(v_n) \\ 0 \le h \le h_{\max}}} \sum x_{ij}^{h,k} \qquad \begin{matrix} \forall v_n \in \mathcal{V} \\ \forall f^k \in \mathscr{F} \end{matrix} \tag{5.11}$$

Moreover, we require that a route, if constructed, must reach all destinations using the following constraint:

$$\sum_{\substack{e_{ij} \in \mathcal{T}(v_n) \\ 0 \le h \le h_{\max}}} \sum x_{ij}^{h,k} \ge \mathbb{1}_{v_n}^{\mathscr{D}^k} \cdot t_{s^k}^k \qquad \begin{matrix} \forall v_n \in \mathcal{V} \\ \forall f^k \in \mathscr{F} \end{matrix} \tag{5.12}$$

Additionally, we must ensure that created routes do not share nodes since the latter are exclusive. In other words, the node can participate in the offloading of at most one flow at a time. This restriction can be stated formally as:

$$\sum_{f^k \in \mathscr{F}} t_n^k \le 1 \qquad \forall v_n \in \mathcal{V} \tag{5.13}$$

$$\sum_{f^k \in \mathscr{F}} \left(\delta_{v_n}^{s^k} + \mathbb{1}_{v_n}^{\mathscr{D}^k} \right) \cdot t_{s^k}^k \le 1 \qquad \forall v_n \in \mathcal{V} \tag{5.14}$$

where the latter constraint deals with the pathological case of a node being both a source and destination (of different flows).

For the half-duplex assignment, the 0-1 variable H_n decides on which half-duplex set, $\mathcal{V}_{\mathrm{H}}^0$ or $\mathcal{V}_{\mathrm{H}}^1$, we put the node v_n. Since a node and its predecessor cannot belong to the same half-duplex set following the alternating link activation strategy, we impose the flowing constraint on the half-duplex assignment:

$$\sum_{\substack{f^k \in \mathscr{F} \\ 0 \le h \le h_{\max}}} \sum x_{ij}^{h,k} \le \mathrm{H}_i + \mathrm{H}_j \le 2 - \sum_{\substack{f^k \in \mathscr{F} \\ 0 \le h \le h_{\max}}} \sum x_{ij}^{h,k} \quad \forall e_{ij} \in \mathscr{E} \tag{5.15}$$

Regarding the RB allocation, we symbolize that v_n be granted the RB (with index) w, for w=1,2,$\cdots$,W, using a 0-1 variable r_n^w. However, since the RB allocations are subjected to the continuity constraint, we employ a matrix-based representation where we enumerate all the feasible RB allocations as columns in a constant 0-1 matrix $\mathcal{Z}_{\mathrm{W}\times\mathrm{U}} = [z_{w,u}]$. The number of columns of this matrix is U=$\frac{\mathrm{W}(\mathrm{W}+1)}{2}$. For instance, the matrix $\mathcal{Z}_{4\times10}$, which enumerates all the feasible contiguous allocations for W=4 RBs, is given below:

$$\mathcal{Z}_{4\times10} = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}$$

where, as an example, we interpret the seventh column as two (contiguous) RB being allocated, namely the 3$^{\mathrm{rd}}$ and the 4$^{\mathrm{th}}$ RBs.

In this matrix formulation, the decision variable r_n^w is linked to another set of auxiliary variables $y_{u,n}$, for u=1,2,$\cdots$,U. Where the variable determines the (unique) column of the matrix $\mathcal{Z}_{\mathrm{W}\times\mathrm{U}}$ that represents the RBs allocated to the node v_n if it acts a (re-)transmitter for some flow. In other words, we have the

following constraint:

$$\sum_{u=1}^{\mathrm{U}} y_{u,n} \leq \sum_{f^k \in \mathscr{F}} t_n^k \qquad \forall v_n \in \mathscr{V} \tag{5.16}$$

The variables $y_{u,n}$ are linked to the variables r_n^w using the elements of the matrix $\mathscr{Z}_{\mathrm{W}\times\mathrm{U}}$ by the following constraint:

$$r_n^w = \sum_{u=1}^{\mathrm{U}} y_{u,n} z_{w,u} \qquad \begin{matrix} \forall v_n \in \mathscr{V} \\ \forall 1 \leq w \leq \mathrm{W} \end{matrix} \tag{5.17}$$

Additionally, the number of RBs allocated to v_n is an integer variable b_n that is linked to the variables r_n^w by the following constraint:

$$b_n = \sum_{w=1}^{\mathrm{W}} r_n^w \qquad \forall v_n \in \mathscr{V} \tag{5.18}$$

Furthermore, using the next constraint, we express the requirement that no relay is allocated more than the RBs requested by the relayed flow.

$$b_n \leq \Omega + (\mathrm{RB}^k - \Omega) \cdot t_n^k \qquad \begin{matrix} \forall v_n \in \mathscr{V} \\ \forall f^k \in \mathscr{F} \end{matrix} \tag{5.19}$$

Using the channel and power model mentioned above, we impose that a RB can be reused for different nodes if harmful interference can be avoided. In other words, we require that the SINR in the system, calculated per RB, must be kept below a predefined threshold γ. We state this formally as the following constraint:

$$\Psi_\sigma r_{ij}^{w,p} + \sum_{v_n \neq v_i} g_{nj} \Psi_{\mathrm{tx}} \cdot \phi_{n,ij}^{w,p} \leq \frac{g_{ij} \Psi_{\mathrm{tx}}}{\gamma} r_{ij}^{w,p} \qquad \begin{matrix} \forall e_{ij} \in \mathscr{E} \\ \forall 1 \leq w \leq \mathrm{W} \\ \forall p \in \{0,1\} \end{matrix} \tag{5.20}$$

where $r_n^{w,p}$ and $r_{ij}^{w,p}$ indicate that the node v_n, the link e_{ij}, respectively, use the RB w during the half-duplex frame-set p. The last constraint also includes an-

other auxiliary 0-1 variable $\phi_{n,ij}^{w,p}$ that indicates that the node v_n interferes with the reception of the link e_{ij}. This variable $\phi_{n,ij}^{w,p} = 1$ only when they are using the RB w and during the same half-duplex frame-set p. The following constraints link the variables $r_n^{w,p}$, $r_{ij}^{w,p}$, and $\phi_{n,ij}^{w,p}$ to the previous variables, along with another link-level 0-1 variable r_{ij} indicates that the link e_{ij} is active in some route:

$$r_n^{w,0} = r_n^w - r_n^{w,1} \qquad \begin{matrix}\forall v_n \in \mathcal{V} \\ \forall 1 \le w \le \mathrm{W}\end{matrix} \tag{5.21}$$

$$r_n^{w,1} = \mathrm{H}_n \cdot r_n^w \qquad \begin{matrix}\forall v_n \in \mathcal{V} \\ \forall 1 \le w \le \mathrm{W}\end{matrix} \tag{5.22}$$

$$r_{ij} = \sum_{\substack{0 \le h \le h_{\max} \\ f^k \in \mathcal{F}}} \sum x_{ij}^{h,k} \qquad \forall e_{ij} \in \mathcal{E} \tag{5.23}$$

$$r_{ij}^{w,p} = r_i^{w,p} \cdot r_{ij} \qquad \begin{matrix}\forall e_{ij} \in \mathcal{E} \\ \forall 1 \le w \le \mathrm{W} \\ \forall p \in \{0,1\}\end{matrix} \tag{5.24}$$

$$\phi_{n,ij}^{w,p} = r_n^{w,p} \cdot r_{ij}^{w,p} \qquad \begin{matrix}\forall v_n \in \mathcal{V}, \forall e_{ij} \in \mathcal{E} \\ \forall 1 \le w \le \mathrm{W} \\ \forall p \in \{0,1\}\end{matrix} \tag{5.25}$$

To keep up the linear formulation, we must linearize the last constraints. To this end, we make use of a standard technique where we introduce an auxiliary 0-1 variable λ_{xy}, for each (binary) product term $x \cdot y$, and three additional linear constraints as follows:

$$(\lambda_{xy} \le x) \wedge (\lambda_{xy} \le y) \wedge (\lambda_{xy} \ge x + y - 1) \tag{5.26}$$

To model the impact of the route $\mathbb{T}^k$, on the energy, we use the empirical model given before to define the following expressions:

$$\Pi_{\mathrm{tx}}^k = \sum_{v_n \in \mathcal{V}} \Pi_{\mathrm{tx},n}^k \cdot t_n^k \tag{5.27}$$

$$\Pi_{\mathrm{rx}}^k = \sum_{\substack{e_{ij} \in \mathcal{E} \\ 0 \le h \le h_{\max}}} \sum \Pi_{\mathrm{rx},ij}^k \cdot x_{ij}^{h,k} \tag{5.28}$$

where Π_{tx}^k and Π_{rx}^k represent the transmitters' and receivers' total power cost,

respectively, for the flow f^k.

In addition to the previous route-based impact on energy, propose to differentiate between nodes according to their residual energy budgets. To this end, we assign each idle node a fractional rank, $\Lambda_n \in (0,1]$, based on the current distribution of residual energy, at the beginning of the frame τ, as follows:

$$\Lambda_n(\tau) = \frac{1}{1+\left\lfloor \frac{\mathrm{E}_n(\tau)-\mathrm{E}_{\min}(\tau)}{\sigma_{\mathrm{E}}(\tau)} \right\rfloor} \tag{5.29}$$

where $\mathrm{E}_{\min}(\tau)$ and $\sigma_{\mathrm{E}}(\tau)$ refer to the minimum and and the standard deviation of residual energy in the network, respectively, at the beginning of frame τ. Note that a higher fractional rank means a higher impact on the node's residual energy.

Finally, we propose to give the ILP model developed so far, an optimality direction by adding the following objective function, which is a normalized equal-weight sum, as follows:

$$\begin{aligned} \max_{x_{ij}^{h,k},\mathrm{H}_n,r_n^w,\cdots} \quad & \frac{1}{\aleph_{\mathrm{B}}}\sum_{v_n\in\mathcal{V}} b_n + \frac{1}{\aleph_{\mathrm{A}}}\sum_{f^k\in\mathscr{F}} t_{s^k}^k - \frac{1}{\aleph_{\mathrm{R}}}\sum_{\substack{v_n\in\mathcal{V}\\ f^k\in\mathscr{F}}}\sum \Lambda_n t_n^k \\ & - \frac{1}{\aleph_{\mathrm{tx}}}\sum_{f^k\in\mathscr{F}} \Pi_{\mathrm{tx}}^k - \frac{1}{\aleph_{\mathrm{rx}}}\sum_{f^k\in\mathscr{F}} \Pi_{\mathrm{rx}}^k \\ \text{subject to:} \quad & (5.6) - (5.25) \end{aligned} \tag{5.30}$$

where we use the next normalizing factors:

$$\begin{aligned} & \aleph_{\mathrm{B}} \triangleq \Omega\cdot|\mathcal{V}|, \aleph_{\mathrm{A}} \triangleq |\mathscr{F}_{\mathrm{C}}|, \aleph_{\mathrm{R}} \triangleq \sum_{v_n\in\mathcal{V}} \Lambda_n, \\ & \aleph_{\mathrm{tx}} \triangleq \sum_{\substack{v_n\in\mathcal{V}\\ f^k\in\mathscr{F}}}\sum \Pi_{\mathrm{tx},n}^k, \aleph_{\mathrm{rx}} \triangleq \sum_{e_{ij}\in\mathscr{E}, f^k\in\mathscr{F}}\sum \Pi_{\mathrm{rx},ij}^k \end{aligned} \tag{5.31}$$

Note that the objective function, as defined above, is a surrogate for the next

eNB's targets to increase: i) the utility of the RB reuse, ii) the number of offloaded flows, and iii) the lifetime of the network and relay nodes.

5.3 Proposals

5.3.1 Exact Resolution Proposal: JRRA-EE

As explained earlier, the eNB must run a centralized algorithm to optimize the overall performance of the offloading scheme developed so far. However, the definition of the optimality target is not straightforward since we have conflicting objectives. On the one hand, for the routing part, one must allocate as few relays as possible to save relays for additional flows and to save energy. On the other hand, one must increase the number of offloaded flows and must consider to longer routes (in terms of relays) to avoid interference zones.

The system can also benefit from optimizing the offloading, considering all the waiting flows as a batch. Such batch treatment adds to the complexity of the problem since it generally implies the combinatorial explosion of the solution space (i.e., the dimension of the latter exponentially increases). As a result, the offloading scheme may not scale well to high-density topologies since it may take too long to optimize the overall problem instance. This complexity can be seen by analyzing the size-complexity of the previous ILP model. It is easy to see that the column-size model (i.e., the number of variables) is, asymptotically, $\mathcal{O}\left(|\mathcal{V}|\mathrm{W}^2+|\mathcal{V}||\mathcal{E}|\mathrm{W}+|\mathcal{V}||\mathcal{F}|+|\mathcal{E}||\mathcal{F}|\,h_{\max}\right)$. Similarly, the row-size of the model (i.e., the number of constraints) is $\mathcal{O}\left(|\mathcal{V}||\mathcal{E}|\mathrm{W}+|\mathcal{V}||\mathcal{F}|+|\mathcal{E}||\mathcal{F}|\,h_{\max}\right)$. This size-complexity adds to the fact that the ILP models are NP-Hard to solve in general.

We propose a two-stage algorithm, named JRRA-EE [41], to solve the ILP model for optimality as it is described above. The first stage in JRRA-EE is pro-

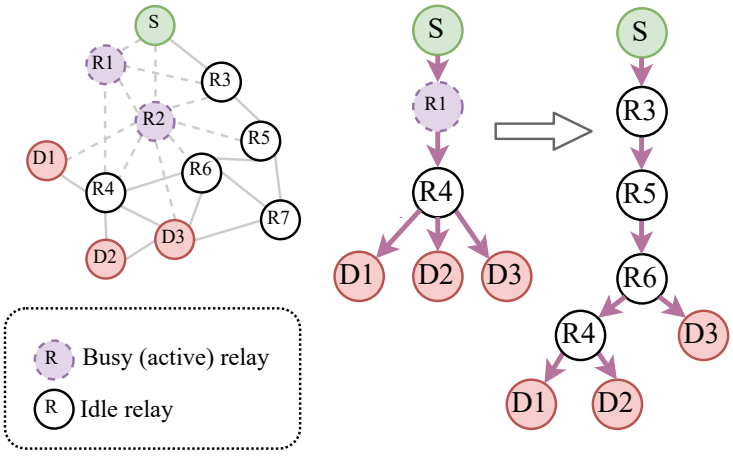


Figure 5.2 – Pre-routing tree formation and its deviation.

posed to decrease the size complexity of the ILP model by reducing the flows $\mathscr{F}_W$ to be admitted in every frame. The initial stage also reduces the model size by finding a reasonable value for the parameter h_{max}. Nevertheless, as seen later, the previous JRRA-EE algorithm does not scale well since it does not converge in a reasonable time for large topologies for the reasons explained above.

JRRA-EE adopts an online bulk strategy by considering, in the SL frame τ resolution, all the scheduled (active) flows, and the waiting flows up to the previous SL frame. However, instead of considering all the waiting flows $\mathscr{F}_W$ for admittance, it proceeds by an initial stage of pre-routing to filter the waiting flows down to a set of candidate flows $\mathscr{F}_C$. The rationale behind this initial stage is to reduce the size complexity of the ILP model by reducing the number of considered flows $\mathscr{F}$ and also by setting the model parameter h_{max} to a reasonable value. It is worth noting that high values for h_{max} implies more possible routing trees to discover while low values few routing trees and hence few admitted flow into the system. The pseudo-code of JRRA-EE illustrated in Algorithm 4.

The pre-routing stage proceeds as follows. For each waiting flow f^k, the eNB checks if it is possible to construct a routing tree from the source node to

Algorithm 4 JRRA-EE pseudo-code

1: **for each** SL frame τ **do**
2: **for each** $f^k \in \mathscr{F}_{\mathrm{A}}$ **do** ▷ Arriving Flows
3: $\mathscr{F}_{\mathrm{W}} \leftarrow \mathscr{F}_{\mathrm{W}} \cup \{f^k\}$
4: **end for**
5: **for each** $f^k \in \mathscr{F}_{\mathrm{FIN}}$ **do** ▷ Finished Flows
6: $\mathscr{V}_{\mathrm{D}} \leftarrow \mathscr{V}_{\mathrm{D}} \cup \mathrm{NodesOF}\left(\mathbb{T}^k\right)$
7: **end for**
8: Execute Algorithm 5 ▷ Pre-routing
9: Construct the ILP model as in formula (3.23)
10: Solve the ILP model using branch-and-cut
11: **for each** $f^k \in \mathscr{F}_{\mathrm{C}}$ **do**
12: **if** $t^k_{s_k} = 1$ **then** ▷ Flow is admitted
13: Configure $\mathbb{T}^k$ according to $x^{h,k}_{ij}$
14: **end if**
15: **end for**
16: $p \leftarrow \tau \mod 2$
17: **for each** $v_n \in \mathscr{V}^p_{\mathrm{G}}$ **do**
18: Allocate RBs for v_n according to $y_{u,n}$
19: **end for**
20: **end for**

all destinations using *breadth-first-traversal* and considering only the currently idle nodes. We recall that each node cannot handle more than one flow. Note that such tree construction stops once all destinations are reached. If such *pre-routing* tree $\widetilde{\mathbb{T}}^k$ exists then the flow f^k is added to the set of candidate flows $\mathscr{F}_{\mathrm{C}}$. In the other case, the flow is kept waiting for upcoming opportunities in subsequent frames. Thanks to the breadth-first-traversal, pre-routing trees are *well-balanced* as they tend to be short one-to-many routing trees. However, due to the dynamic state (e.g., end of current flows, low battery, etc.) of nodes, pre-routing trees also tend to deviate from this preferred condition, as illustrated in Fig. 5.2. The pseudo-code of the pre-routing tree's construction is illustrated in Algorithm 5.

Taking advantage of the dynamic nature of pre-routing trees construction

Algorithm 5 Pre-routing of routing trees pseudo-code

Inputs: $\mathcal{V}_\mathrm{D}, \mathbb{T}^k \,\forall f^k \in \mathscr{F}_\mathrm{S}$, $\mathscr{F}_\mathrm{W}$

Outputs: $\mathscr{F}_\mathrm{C}, h_\mathrm{max}$

1: $\mathscr{F}_\mathrm{C} \leftarrow \emptyset, h \leftarrow 0, \widetilde{h} \leftarrow 0$
2: **for each** $f^k \in \mathscr{F}_\mathrm{S}$ **do** ▷ Trees of active flows
3: **if** HeightOF$\left(\mathbb{T}^k\right) > h$ **then**
4: $h \leftarrow$ HeightOF$\left(\mathbb{T}^k\right)$
5: **end if**
6: **end for**
7: **for each** $f^k \in \mathscr{F}_\mathrm{W}$ **do**
8: **if** $\{v_{s^k}\} \cup \mathscr{D}^k \nsubseteq \mathcal{V}_\mathrm{D}$ **then go to** 30
9: **end if**
10: $\mathrm{Q} \leftarrow \emptyset$ ▷ New empty queue
11: push v_{s^k} into Q
12: $\mathrm{S} \leftarrow \{v_{s^k}\}$
13: LevelOF$(v_{s^k}) \leftarrow 0$
14: **while** $\mathrm{Q} \neq \emptyset \wedge \mathscr{D}^k \nsubseteq \mathrm{S}$ **do** ▷ Breadth-first traversal
15: $v_i \leftarrow$ Q.pop()
16: **if** LevelOF$(v_i) > \widetilde{h}$ **then**
17: $\widetilde{h} \leftarrow$ LevelOF(v_i)
18: **end if**
19: **for each** $e_{ij} \in \mathscr{O}(v_i)$ **do**
20: **if** $v_j \notin \mathrm{S} \wedge v_j \in \mathcal{V}_\mathrm{D}$ **then**
21: push v_j into Q
22: $\mathrm{S} \leftarrow \mathrm{S} \cup \{v_j\}$
23: LevelOF$(v_j) \leftarrow$ LevelOF$(v_i) + 1$
24: **end if**
25: **end for**
26: **end while**
27: **if** $\mathscr{D}^k \subseteq \mathrm{S}$ **then** ▷ Add f^k to candidates
28: $\mathscr{F}_\mathrm{C} \leftarrow \mathscr{F}_\mathrm{C} \cup \{f^k\}$
29: **end if**
30: **end for**
31: $h_\mathrm{max} \leftarrow \max\{h, \beta \cdot \widetilde{h}\}$ ▷ Update h_max

stage, the latter goes one step further to set the parameter value h_max of the current ILP model based on the reported trees heights and those of the routing

trees of active scheduled flows as follows:

$$h_{\max} = \max\left\{ \max_{f^k \in \mathscr{F}_S} \mathfrak{H}\left(\mathbb{T}^k\right), \beta \max_{f^k \in \mathscr{F}_C} \mathfrak{H}\left(\widetilde{\mathbb{T}}^k\right) \right\} \tag{5.32}$$

where i) $\mathfrak{H}(\cdot)$ denotes the height-of-tree operator and ii) $\beta \geq 1$ is a "tradeoff-margin" factor to allow for longer routing trees to be explored and more flows to be admitted into the system when solving the current ILP model. After this initial stage, we solve the reduced ILP model using the celebrated Branch-and-Cut method.

5.3.2 Novel Heuristic-Based Proposal: HERRA

To address the scalability issue in JRRA-EE as discussed above, we employ a new heuristic-based method, called Heuristic Energy-aware Routing and RB Allocation (HERRA), to reduce the complexity of the problem. We propose to decompose the resolution of the problem into three stages: i) the iterative energy-aware routing, ii) the flow conflict resolution, and iii) the RB allocation.

In the first stage, the iterative routing attempt to find the routes for the waiting flows, one after the other, regardless of their competition for relay nodes. Then, the second stage intervenes to resolve the conflicts between the flows according to some strategy. Lastly, an algorithm to do the RB allocation follows. It is straightforward to see that this significantly reduces the (time) complexity at the expense of optimality. This scheme dispenses with the batch and joint treatment with the implied intricacies in optimizing the decision over all the feasible route combinations of flows and all possible RB allocations at the same time.

For the first stage of our scheme, we propose an algorithm based on Dijkstra's algorithm to minimize the number of relay nodes in a route together with

the route impact on the residual energy budgets. We propose to assign each (idle) node in the topology a node cost (weight) as follows:

$$c_n(\tau) = \theta + \exp\left(-\frac{\mathrm{E}_n(\tau)}{\mathrm{E}_n(0)-\mathrm{E}_n(\tau)}\right) \tag{5.33}$$

where θ is a non-negative parameter that balances between minimizing the number of nodes, which leaves more relays available for the next flow, and minimizing the impact on energy.

Note that exponential part in (5.33) starts as a cost of 0 for $\exp(-\infty)$ and tends to 1 as the node's energy budget depletes. For this reason, plausible values of θ in (5.33) should be in the interval $[0,1]$. For higher values, $\theta > 1$, the routing will tend towards minimizing the number of nodes in the route. In the routing algorithm, we turn the node costs into edge costs as follows. As we add edges to the route (tree), the cost of an edge will be the cost of the other endpoint not already part of the route.

For each flow f^k in $\mathscr{F}_\mathrm{C} \subseteq \mathscr{F}_\mathrm{W}$, the routing algorithm starts, for the first destination, as the usual Dijkstra's algorithm with the cost structure defined above. However, the cost structure is adjusted for the subsequent iterations for the other destinations. Every time, the algorithm finds a path (for the destination), it updates the evolving route with the new edges. The cost structure is updated in the following way. If an edge is already part of $\mathbb{T}^k$, it gets a cost of 0 to encourage the system to reuse the already selected relays. Otherwise, if an edge arrives into a node in $\mathbb{T}^k$, the edge is skipped. In other words, no node receives more than a unique incoming edge. When this procedure ends, if all destinations are reached, the flow f^k is added to the newly admitted flow $\mathscr{F}_\mathrm{N}$. The pseudo-code for this routing stage is given in Algorithm 6.

However, the first stage can produce routes, for the newly admitted flows, which overlap (i.e., share nodes among them). Therefore, the second stage re-

Algorithm 6 Pseudo-code for HERRA: the routing algorithm

Inputs: $\mathcal{V}_\mathrm{D}, \mathscr{F}_\mathrm{C}, \{c_n,\ \forall v_n \in \mathcal{V}_\mathrm{D}\}$

Outputs: $\mathscr{F}_\mathrm{N}, \{\mathbb{T}^k,\ \forall f^k \in \mathscr{F}_\mathrm{N}\}$

1: $\mathscr{F}_\mathrm{N} \leftarrow \emptyset$
2: **for each** $f^k \in \mathscr{F}_\mathrm{C}$ **do**
3: $\mathbb{T}^k \leftarrow \emptyset$
4: **for each** $d^k \in \mathscr{D}^k$ **do**
5: **if** $d^k \in \mathbb{T}^k$ **then go to** 4 **end if**
6: **for each** $v_n \in \mathcal{V}_\mathrm{D}$ **do**
7: $\mathrm{Q} \leftarrow \mathrm{Q} \cup \{v_n\}$, $\mathrm{dist}[v_n] \leftarrow \infty$, $\mathrm{pred}[v_n] \leftarrow$ **nil**
8: **end for**
9: $\mathrm{dist}[s^k] \leftarrow 0$
10: **while** $\mathrm{Q} \neq \emptyset$ **do**
11: $v_n \leftarrow \mathrm{argmin}_{v_x \in \mathrm{Q}} \mathrm{dist}[v_x]$, $\mathrm{Q} \leftarrow \mathrm{Q} \setminus \{v_n\}$,
12: **if** $v_n = d^k$ **then go to** 25 **end if**
13: **for each** $e_{nj} \in \mathscr{O}(v_n) \mid v_j \in \mathrm{Q}$ **do**
14: **if** $v_j \notin \mathbb{T}^k$ **then**
15: $\mathrm{newdist} \leftarrow \mathrm{dist}[v_n] + c_j$
16: **else if** $v_n \in \mathbb{T}^k$ **then**
17: $\mathrm{newdist} \leftarrow \mathrm{dist}[v_n]$
18: **else go to** 13
19: **end if**
20: **if** $\mathrm{newdist} < \mathrm{dist}[v_j]$ **then**
21: $\mathrm{dist}[v_j] \leftarrow \mathrm{newdist}$, $\mathrm{pred}[v_j] \leftarrow e_{nj}$
22: **end if**
23: **end for**
24: **end while**
25: **if** $\mathrm{pred}[d^k] \neq$ **nil then**
26: $v_n \leftarrow d^k$
27: **while** $v_n \neq s^k$ **do**
28: $e_{in} \leftarrow \mathrm{pred}[v_n]$, $\mathbb{T}^k \leftarrow \mathbb{T}^k \cup \{e_{in}\}$, $v_n \leftarrow v_i$
29: **end while**
30: **else go to** 2
31: **end if**
32: **end for**
33: $\mathscr{F}_\mathrm{N} \leftarrow \mathscr{F}_\mathrm{N} \cup \{f^k\}$
34: **end for**

solves these flow conflicts according to some strategy by rejecting some flows and putting them back to the waiting queue. To record the conflicts, we employ a conflict graph $\mathbb{F}$. In $\mathbb{F}$, we represent a flow by a node where an edge between two flows means that the flows conflict with each other (i.e., they share some relay nodes). Formally, this flow conflict graph $\mathbb{F}$ is defined as:

$$\mathbb{F} = \left\{ (f^{k'}, f^{k''}) \mid f^{k'}, f^{k''} \in \mathscr{F}_{\mathrm{N}}, \mathbb{T}^{k'} \cap \mathbb{T}^{k''} \neq \emptyset \right\} \tag{5.34}$$

To continue resolution, we propose to go over the flows in $\mathbb{F}$ in an order defined by a strategy. The first strategy is a First-Come First-Served (FCFS) strategy. Another strategy is to iterate, starting with the Least-Conflicting Flows First (LCFF). In either case, every time we accept a flow f^k from $\mathbb{F}$, we remove its conflicting flows from $\mathbb{F}$ (i.e., those having edges with f^k). This process continues until we process the whole $\mathbb{F}$. This procedure reduces the newly admitted flows $\mathscr{F}_{\mathrm{N}}$ by removing the conflicts.

The last stage of the scheme HERRA is to allocate RBs for the newly admitted flows. To this end, we propose two strategies for this step. The first strategy is the Random RB allocation (RRA). In this strategy, for each new flow f^k, iteratively, we randomly assign a half-duplex set for the source node (root) of the $\mathbb{T}^k$. In doing so, the half-duplex set assignment is done for the other nodes in $\mathbb{T}^k$ according to the principle of alternating link scheduling explained earlier. And then, we randomly allocate a feasible RB pattern (i.e., a column of the matrix $\mathscr{Z}_{\mathrm{W}\times\mathrm{U}}$) for each node in $\mathbb{T}^k$. However, only those columns, which allocate no more RBs than the request of the flow, RB^k, are considered with equal probabilities.

In addition to RRA, we propose another interference-aware RB allocation strategy (IRA). In IRA, the half-duplex assignment proceeds as before. However, for the RB allocation, we iteratively build a small ILP model for each flow.

This ILP seeks to minimize the interference with the already-scheduled flows due to the current RB allocation. To consider interferences, we define interference budgets for the already schedule nodes before the RB allocations of the current flow. These budgets are defined in a manner compatible with the SINR threshold γ. This interference budget η_n^w for the node $v_n \in \mathcal{V}_{\mathrm{H}}^p$, over the RB w, is calculated as:

$$\eta_n^w = \begin{cases} +\infty & \text{if } r_n^w = 0 \\ \frac{g_{mn}\Psi_{\mathrm{tx}}}{\gamma} - \Psi_\sigma - \sum\limits_{v_i \in \mathcal{V}_{\mathrm{H}}^p \setminus \{v_n\}} g_{in}\Psi_{\mathrm{tx}} \cdot r_i^w & \text{if } r_n^w = 1 \end{cases}$$

The proposed ILP model includes the 0-1 variable r_n^w as defined before for the nodes in the current $\mathbb{T}^k$. However, to formulate the contiguity constraint, we dispense with the matrix formulation and adopt the graph-based one presented in Chapter 4. The latter gives a more tractable ILP, and hence more rapid to solve, than the matrix-based one. Recall that in this method, we represent a RB, being allocated, by a selection of an arc in a RB Allocation Graph (RBAG), as illustrated in Figure 4.1. As illustrated in Section 4.2, in the RBAG, we replace each RB w by a vertex and add an arc between each RB w and its successor w+1. Besides, in the RBAG, we add a virtual source and sink vertices. Also, we include arcs from the virtual source to the RBs to encode the start position of an allocation of RBs. Similarly, we add arcs from each RB to the virtual sink to encode where the allocation ends. We note that every contiguous allocation matches a (continuous) path on the RBAG starting from the source vertex and ending into the sink vertex.

For the ILP formulation, we have additional 0-1 variables y_n^w and z_n^w, which correspond to the arcs from, and into, virtual vertices respectively.

The ILP model of `IRA`, for each $f^k \in \mathscr{F}_{\mathrm{N}}$, is defined in Model 1. At the end of

the ILP solution, if a solution is found, the IRA adds the flow f^k to the scheduled flows $\mathscr{F}_{\mathrm{S}}$ and updates the sets, $\mathcal{V}_{\mathrm{H}}^0, \mathcal{V}_{\mathrm{H}}^1$ and the interference budgets, η_n^w, accordingly before proceeding with the next flow. Otherwise, the flow f^k is rejected and is put back into the waiting queue.

Model 1 The ILP model of IRA for the flow $f^k \in \mathscr{F}_{\mathrm{N}}$.

$$\min_{r_n^w, y_n^w, z_n^w \mid v_n \in \mathbb{T}^k} \sum_{\substack{1 \le w \le \mathrm{W} \\ p \in \{0,1\} \\ e_{ij} \in \mathcal{V}_{\mathrm{H}}^p}} \sum \sum \left\{ \sum_{e_{mn} \in \mathbb{T}^k \mid \mathrm{H}_m = p} g_{in} \Psi_{\mathrm{tx}} \cdot r_m^w \right\} \tag{5.35}$$

subject to:

$$\sum_{e_{mn} \in \mathbb{T}^k \mid \mathrm{H}_m = p} g_{in} \Psi_{\mathrm{tx}} \cdot r_m^w \le \eta_j^w \qquad \begin{array}{l} \forall 1 \le w \le \mathrm{W} \\ \forall p \in \{0,1\} \\ \forall e_{ij} \in \mathcal{V}_{\mathrm{H}}^p \end{array}$$

$$y_n^w + r_n^{w-1} = r_n^w + z_n^w \qquad \begin{array}{l} \forall v_n \in \mathbb{T}^k \\ \forall 1 \le w \le \mathrm{W}+1 \end{array}$$

$$\sum_{w=1}^{\mathrm{W}+1} r_n^w = b \qquad \forall v_n \in \mathbb{T}^k$$

$$\sum_{w=1}^{\mathrm{W}+1} y_n^w = 1 \qquad \forall v_n \in \mathbb{T}^k$$

$$\sum_{w=1}^{\mathrm{W}+1} z_n^w = 1 \qquad \forall v_n \in \mathbb{T}^k$$

$$1 \le b \le \mathrm{RB}^k$$

5.4 Performance Evaluation

This section give the performance evaluation of our proposed scheme HERRA. As we did in Chapter 3 and Chapter 4, we base our evaluation on extensive network simulation in our extended NS-3 environment that includes a support for LTE-D2D protocol stack as described in Section 3.4. We also define our evaluation metrics before we describe, in detail, the executed simulation scenarios.

5.4.1 General Scenario Parameters

In our experiments, we employ the same parameters of the the scenario in Section 3.4. Same as before, we simulate a cellular network of a single non-sectorized cell with a radius of R_{cell} = 1 km, which is managed by one LTE-A eNB. Unless stated otherwise, we use the same parameter values given in Table 3.1. Besides, for the density of the nodes, λ_{UE}, distributed as a Poisson Point Process (PPP), we use values in the range [10−80] nodes per km^2. Moreover, we assume that every node starts with an initial energy budget $E_n(0)$=3.856 Joules.

5.4.2 Simulated Traffic Parameters

Similar to Chapter 3, to simulate the data traffic, we use a Poisson arrival process to generate multicast traffic flows. To produce different traffic load condition, we set the arrival rate of flows λ_{FL} to values from $\{10, 20\}$ flows per second. Additionally, we suppose that the flows have different Constant Bit-Rate (CBR) selected randomly from the predefined classes shown in Table 3.1. As for the duration of flows, we assume that the values follow an exponential random variable with a mean duration of λ_{DUR} = 1 second. For the selection of the sources and destinations, we use the *fixed subscription-rate* as Section 4.4 to model the probability of a node to subscribe as a receiver. To reiterate it here, we select the sources randomly using a uniform distribution and, then, for destinations, we assume a node-flow interest probability ρ=0.1, where we decide whether every other node is a receiver using Bernoulli trials with a success probability of ρ.

5.4.3 Performance Metrics

Let $\mathscr{F}_{\text{TOT}} \supseteq \mathscr{F}_{\text{ADM}} \supseteq \mathscr{F}_{\text{INT}}$ be the total sets of arrived, admitted and interrupted flows, respectively, during a simulation run. Also, let $\mathrm{E}[\cdot]$ denote the average sample metric over all the simulation runs. Then, to evaluate the performance of the algorithm `HERRA` versus the exact resolution method `JRRA-EE`, we retain the metrics, $\mathbb{S}$, $\mathbb{H}$, and $\mathbb{L}$ from Section 3.4.3 in addition to the following metrics:

- The metric $\mathbb{L}$ is modified to take into account the multicast case and the interrupted flows as follows:

$$\mathbb{L} = \mathrm{E}\left[\frac{\sum\limits_{f^k \in \mathscr{F}_{\text{ADM}} \setminus \mathscr{F}_{\text{INT}}} \frac{|\mathscr{D}^k|\mathtt{pkts}_{\text{tx}}^k - \mathtt{pkts}_{\text{rx}}^k}{|\mathscr{D}^k|\mathtt{pkts}_{\text{tx}}^k}}{|\mathscr{F}_{\text{ADM}} \setminus \mathscr{F}_{\text{INT}}|}\right] \tag{5.36}$$

- The average ratio of the interrupted flows due to node-exits after exceeding the allocated energy budget. Formally, this metric $\mathbb{I}$ is defined as:

$$\mathbb{I} = \mathrm{E}\left[\frac{|\mathscr{F}_{\text{INT}}|}{|\mathscr{F}_{\text{ADM}}|}\right] \tag{5.37}$$

- The mean occurrence of node-exit events due to the energy budget limitation. Formally, we define this metric $\mathbb{E}$ as follows:

$$\mathbb{E} = \mathrm{E}\left[\frac{|\mathscr{V}_{\text{X}}|}{\texttt{total duration of the simulation run}}\right] \tag{5.38}$$

- The average computation time required to solve one occurrence of the whole routing and RB allocation problem during a frame. This metric $\mathbb{C}$

is defined by:

$$\mathbb{C} = \mathrm{E}\left[\frac{\sum_{\tau} \texttt{solution time of the frame } \tau}{\texttt{total frames in the simulation run}}\right] \tag{5.39}$$

5.4.4 Simulation Results

For the initial evaluation of our scheme HERRA, we will use a basic variant FCFS with the random RB allocations (i.e., HERRA+FCFS+RRA). We set the parameter theta to values from the range [0−1] with a step of 0.25. We compare the performance of HERRA to that of the former JRRA-EE. The results of JRRA-EE are presented partially (i.e., for $\lambda_{\mathrm{UE}} \in [10-40]$) because of the non-scalability of JRRA-EE as it expends tremendous time to solve the implied model to optimality as shown later.

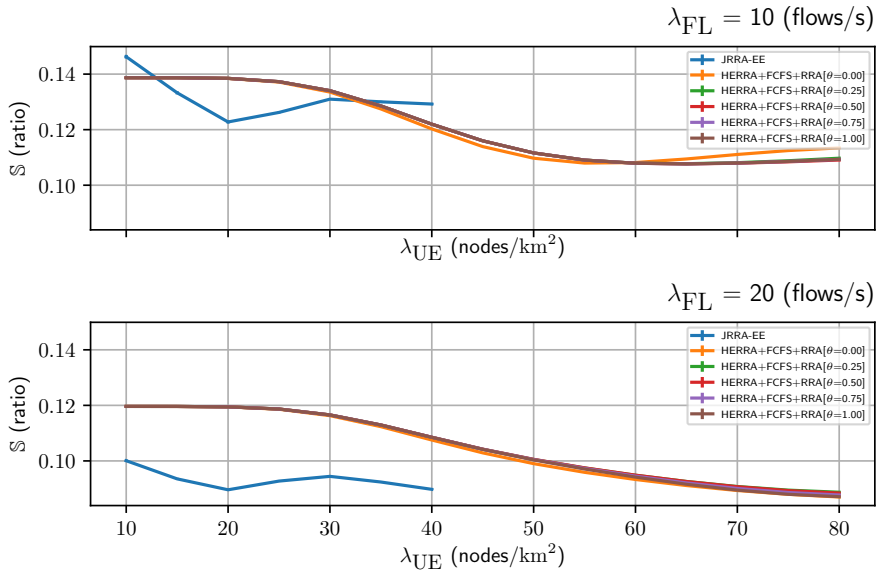


Figure 5.3 – Performance of HERRA with respect to the offloading ratio $\mathbb{S}$.

Figure 5.3 shows the plot of the metric $\mathbb{S}$ for the basic scheme of HERRA accompanying the scheme JRRA-EE. Under the two traffic conditions, we observe that HERRA generally accepts more flows to offload than JRRA-EE. This

advantage is very noticeable in high traffic load conditions where the difference can reach a $\Delta\mathbb{S}$=0.02. Also, we notice that this offloading ratio decreases with the density of the topology and the intensity of the traffic represented by λ_{UE} and λ_{FL} respectively. We can explain this by the fact that JRRA-EE is more conservative in admitting flows into the system since it must strictly check the interference-tolerance condition.

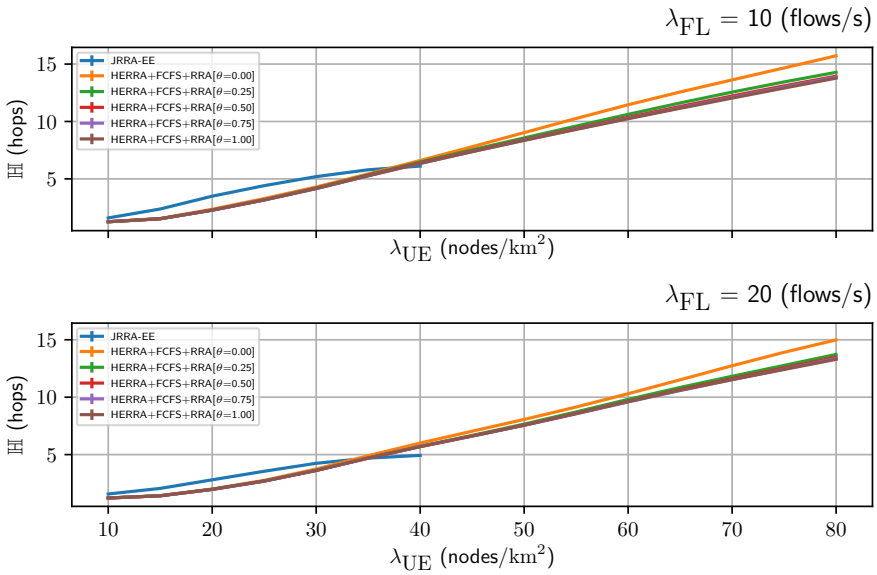


Figure 5.4 – Performance of HERRA with respect to the hop count $\mathbb{H}$.

Figure 5.4 shows the performance in terms of the metric $\mathbb{H}$, which indicates the hop count in the routes. This metric also gives the QoS offered to the flows in the matter of latency where low values of $\mathbb{H}$ means little delay. Concretely, the end-to-end and the average packet delays are in proportion to the product $\mathbb{H} \cdot T_{SL}$. The figure also reveals that the hop count (i.e., the tree height of the route) increases approximately in linear relation to the density of nodes λ_{UE}. Moreover, the plots reveal that the basic variant of HERRA has an advantage over JRRA-EE for low-density topologies. We note that higher values of θ enhance the performance by yielding shorter routes.

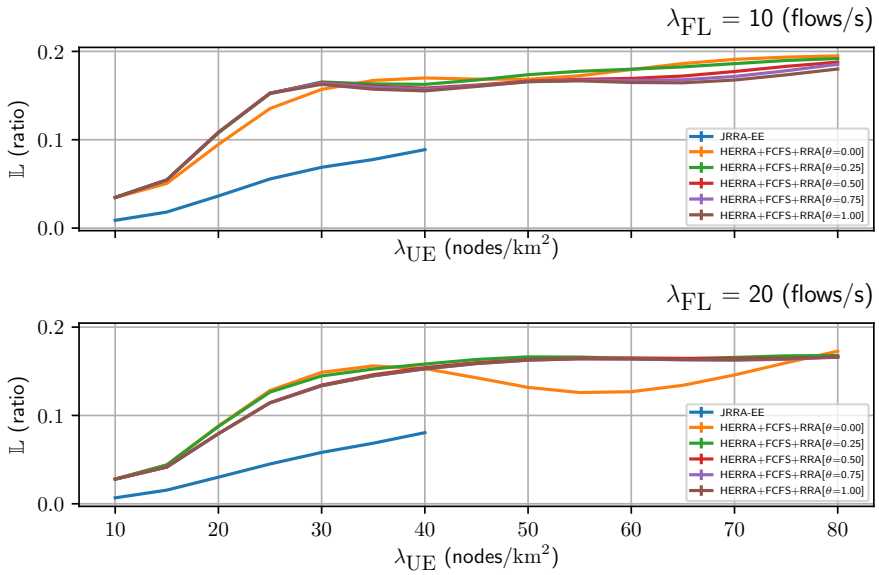


Figure 5.5 – Performance of HERRA with respect to the packet loss $\mathbb{L}$.

The packet loss metric $\mathbb{L}$, which is another QoS-related metric, is shown in Figure 5.5. The plots reveal that the JRRA-EE gives a superior performance than that of HERRA. This relative advantage of JRRA-EE is a consequence of being strict in keeping interferences below the approved level. However, we note that the selected variant of HERRA strives to minimize the packet loss with the increased node density λ_{UE} under the presented traffic load conditions.

The mean ratio of interrupted flows in the system is given by the plots of the metrics $\mathbb{I}$ in Figure 5.6. This metric reflects two important aspects of the system. First, from the viewpoint of flows, this quantifies the service continuity, which can be considered a QoS metric. An interrupted flow means that the system must revert back to the conventional cellular method to ensure the service continuity. Secondly, and most importantly, this metric indicates the degree of energy-awareness of the whole system. High interrupt rates mean that the system fails to harvest the available energy budget to the benefit of the offloading service and to increase its utility. From the plots of $\mathbb{I}$ in the Figure 5.6, we easily

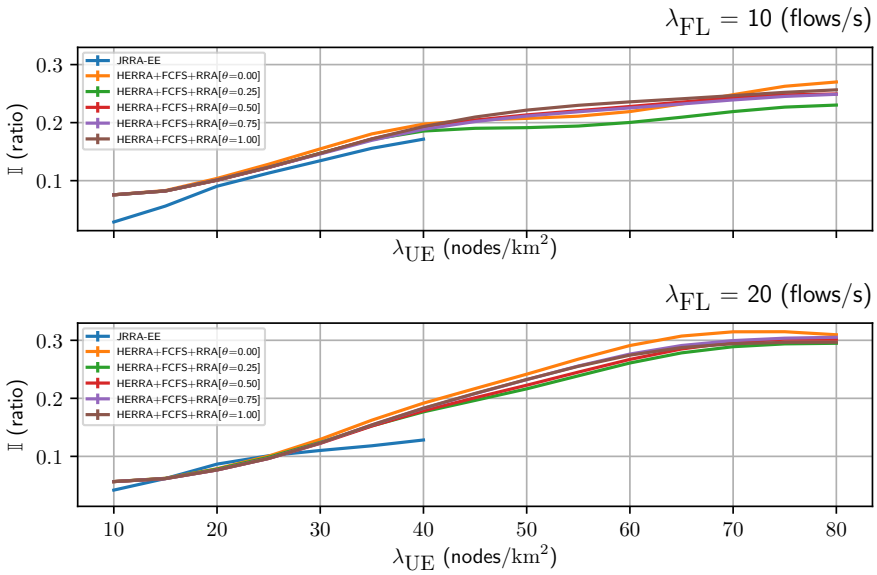


Figure 5.6 – Performance of HERRA with respect to the flow interruption ratio $\mathbb{I}$.

see that the original non-scalable scheme JRRA-EE outperforms the selected variants of HERRA for low-density topologies. Nevertheless, the latter performs very well especially for $\lambda_{UE} \leq 30$ and it strives to keep the interruption rate below 0.3 for denser topologies. We also observe an advantage for the variant with θ=0.25 under the two traffic load conditions.

In the same vein, the occurrence rate of the node-exit events due to energy depletion is estimated using the metric $\mathbb{E}$ whose plots are illustrated in Figure 5.7. This metric is the node-level counterpart of $\mathbb{I}$, and similarly, it also reflects the energy-awareness of the used algorithm. The evolution of $\mathbb{E}$ confirms the same conclusions of $\mathbb{I}$ about the advantage of JRRA-EE over HERRA with the relative advantage of the variant with θ=0.25.

On the Effect of Flow Conflict Resolution Strategy

To compare the two strategies, FCFS and LCFF, used to resolve flow conflicts in the second stage of the proposed HERRA, we fix the value of θ used in the routing

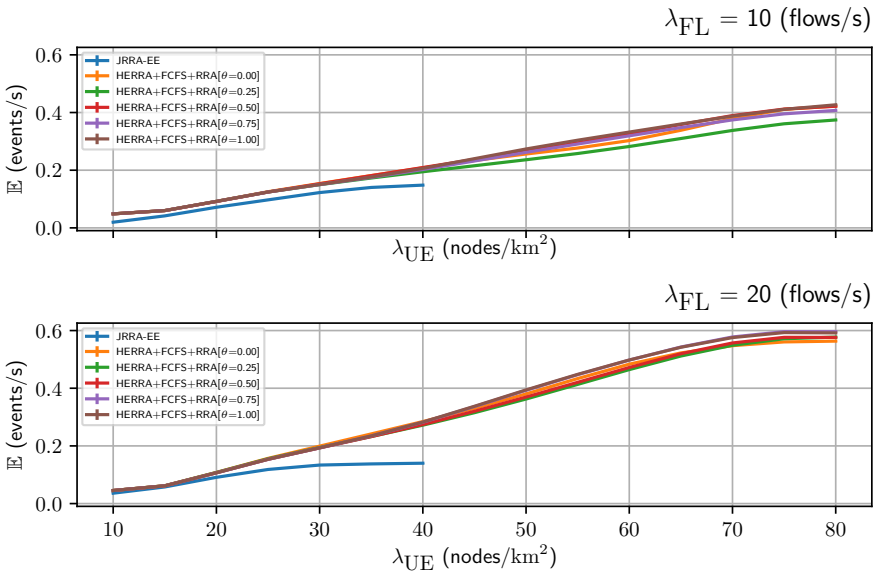


Figure 5.7 – Performance of HERRA with respect to the occurrence rate of the interruption event (node-exit) $\mathbb{E}$.

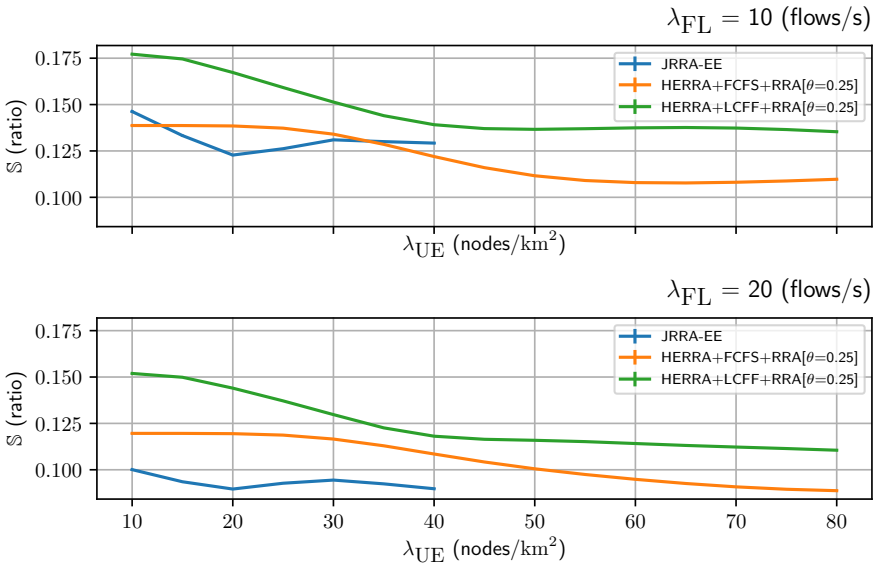


Figure 5.8 – Comparison between the strategies FCFS and LCFF with respect to the offloading ratio $\mathbb{S}$.

stage to θ=0.25. We continue to use the basic Random RB Allocation RRA. The comparative performance of these two HERRA variants, HERRA+FCFS+RRA and

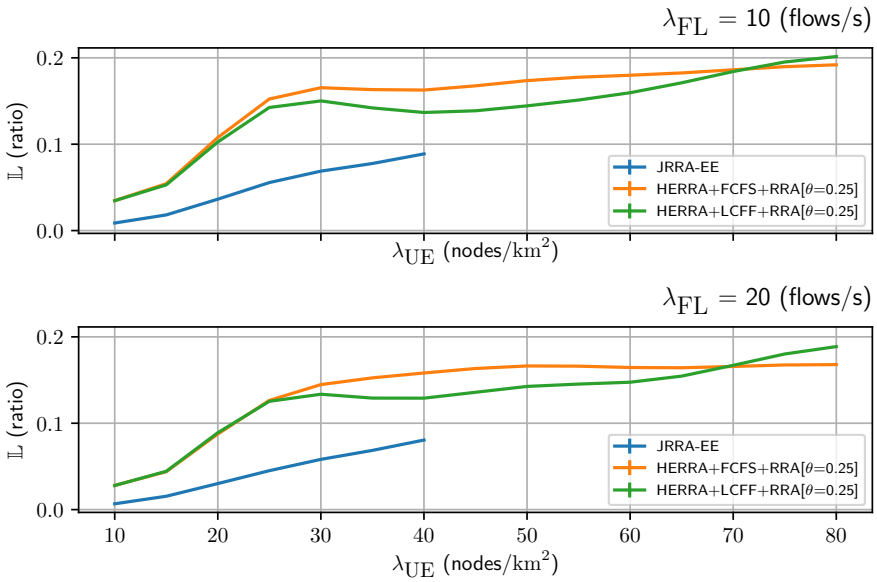


Figure 5.9 – Comparison between the strategies FCFS and LCFF with respect to the packet loss ratio $\mathbb{L}$.

HERRA+LCFF+RRA, are given in Figure 5.8 and Figure 5.9, as regards the metrics $\mathbb{S}$ and $\mathbb{L}$. Figure 5.8 proves the advantage of LCFF over the basic FCFS in terms of the offloading ratio. Recall that LCFF works by accepting the flows with least conflicts in each frame, and hence it is likely to take in more flows than the simple FCFS strategy. Moreover, Figure 5.9 shows that the variant LCFF outperforms FCFS up to a certain node-density in terms of the packet loss.

On Effect of RB Allocation Method

Using the same parameter value θ=0.25, we compare the basic scheme HERRA+FCFS+RRA to its counterpart HERRA+FCFS+IRA. In the latter, we employ the interference aware strategy IRA instead of the random RRA. Figure 5.10 and Figure 5.11 highlight that IRA give more reliable service than RRA, in terms of $\mathbb{L}$ as shown in Figure 5.10, at the expense of low acceptance ratio $\mathbb{S}$ as revealed in Figure 5.10. Recall, that because the IRA variant is interference aware, com-

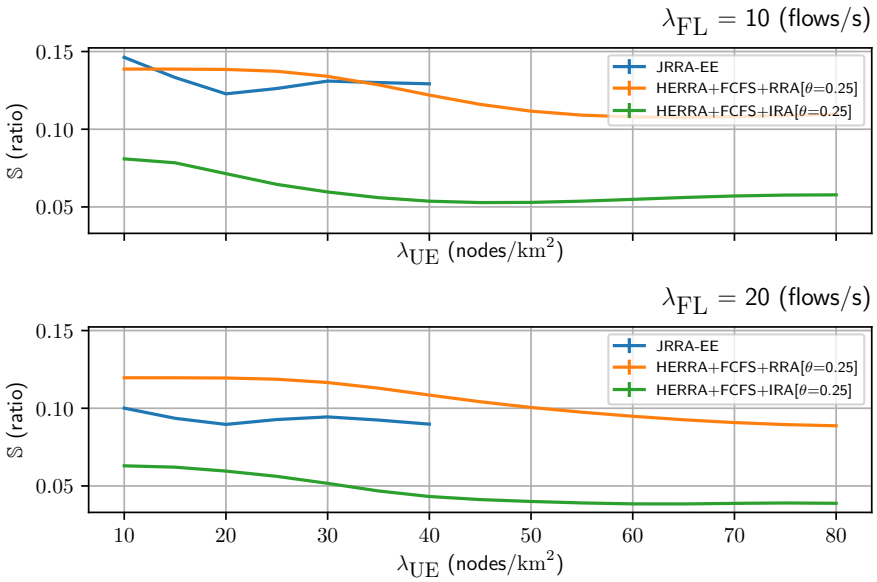


Figure 5.10 – Comparison between the strategies RRA and IRA with respect to the offloading ratio $\mathbb{S}$.

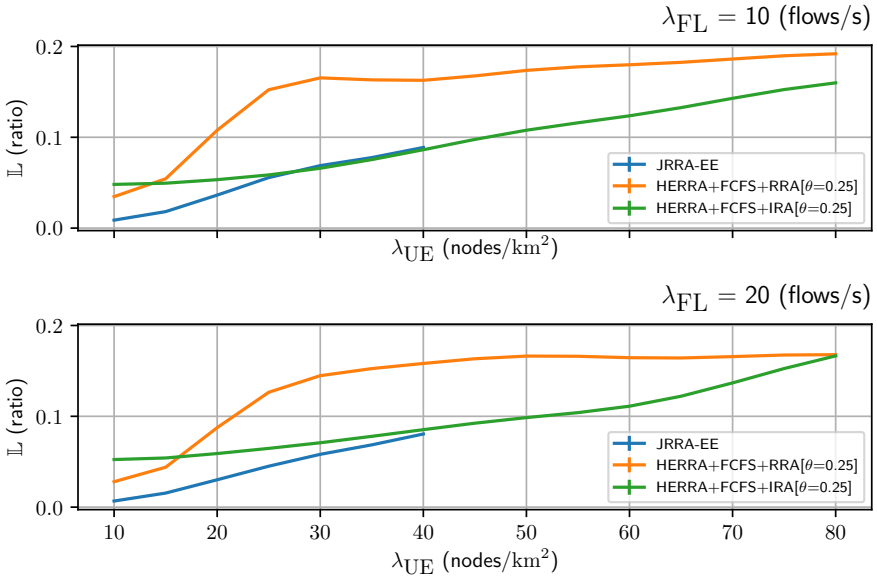


Figure 5.11 – Comparison between the strategies RRA and IRA with respect to the packet loss ratio $\mathbb{L}$.

pared to RRA, the IRA can reject a flow accepted by the previous two stages of HERRA if the RB allocation cannot be done respecting the interference limit.

On the Computation Time and Scalability

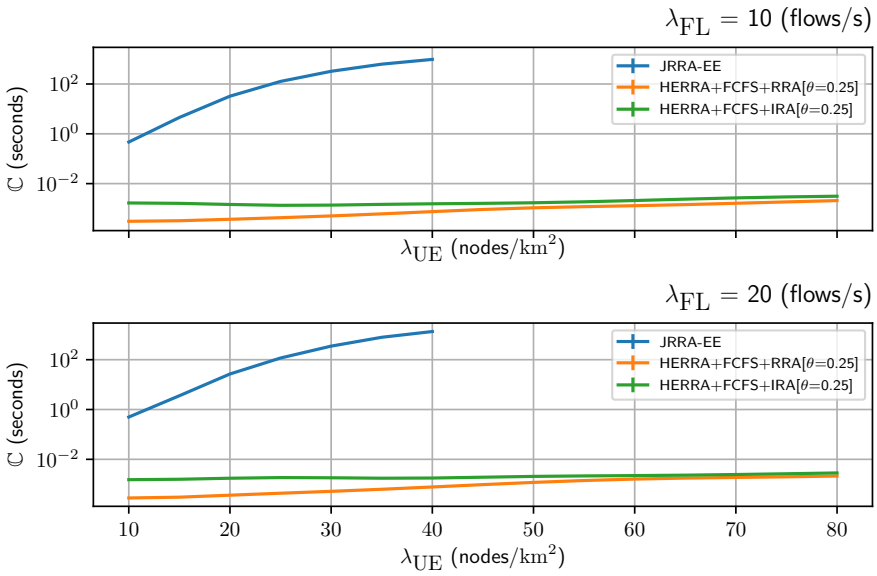


Figure 5.12 – Comparison between the offloading schemes time with regards to the computation time $\mathbb{C}$.

To highlight the scalability of the proposal HERRA in relation to the original JRRA-EE scheme, we present the computation time metric $\mathbb{C}$ in Figure 5.12. For comparison, we choose the variants: HERRA+FCFS+RRA and HERRA+FCFS+IRA with θ=0.25. The evolution of $\mathbb{C}$ in relation to the node-density, λ_{UE}, shows the reason behind the non-scalability JRRA-EE. Indeed, the JRRA-EE may take up to about 1000 seconds, for λ_{UE}=40, to solve its combinatorially-complex ILP-model to optimality. However, for a practical scheme, the total solution time must be on the timescale of the frame, T_{SL}, which can be as low as 40 ms as in our simulations. On the other hands, the proposal HERRA, as shown in Figure 5.12, scales well to denser topologies and yields solutions on the order of 1 ms. In other words, HERRA can achieve massive speedups, relative to JRRA-EE, up to six orders of magnitude. We notice that the variant IRA takes longer to solve than the basic RRA because the former incorporates non-complex ILP

models in its solution.

To sum up the results, the proposal HERRA offers a more scalable, and largely more practical, offloading service than the original optimal JRRA-EE at the cost of less reliable service. We can obtain further enhancements by choosing one of the variants of HERRA. Moreover, the algorithm can be tweaked further, to reach performance goals, by varying the parameter θ, which is beyond the scope of this chapter.

5.5 Conclusion

In this chapter, we considered the design of energy-aware traffic offloading schemes based on LTE-D2D for UE-to-UEs traffic in LTE-A cellular systems. Precisely, the concept was about a single LTE-A eNB being able to relieve itself from routing UE-to-UEs data flows by exploiting a collaborating relaying network of D2D-capable UEs. Moreover, the eNB controls the routing and the allocation of the OFDMA RBs during this operation. We presented two schemes to solve the routing and RB allocation. Both methods recognize the essential details of LTE-D2D: namely the contiguous RB allocation and the half-duplex nature of D2D. In addition to energy-awareness, the first scheme is an optimal one that solves the problems of the routing and RB allocation jointly based on a non-scalable ILP formulation. To address this scalability issue, we presented a novel scheme to solve the problem using a more tractable heuristic algorithm. The heuristic scheme is a parametric three-stage method which includes variations on the strategies used in its stages. Our evaluation based on network simulations, using NS-3, proved that our new proposal converges rapidly yielding massive speedups. Therefore, performance evaluation showed also that the new scheme is more scalable and more practical than the original one. Nev-

ertheless, due to speed-optimality tradeoffs, the new scheme has small performance gaps relative to the original one, especially in the matter of the reliability of the service.

Chapter 6

Conclusion

"We must, after all, leave something for the future."

— Richard Feynman, *The Feynman Lectures on Physics, Vol. III*

6.1 Summary of Contributions

Throughout this thesis, we addressed the problem of solving routing and RB allocations in multihop LTE-D2D communications within LTE-A cellular systems. We focused on designing D2D-based offloading schemes where the problems mentioned above are solved by the eNB, which acts as a centralized controller. Also, we paid attention to the practical aspect of LTE-D2D to ensure that our proposed scheme is feasible. Namely, we considered the allocation of frequency resources in terms of RBs, according to traffic requirements, while guaranteeing their contiguity. Besides, we took into account the half-duplex mode of operations in UEs. The traffic type we considered to offload is UE-to-UE. Moreover, we validated our proposals using the NS-3 network simulators, which we had extended to support LTE-D2D. Besides all these commonalities, each presented contribution has its specific scope as follows:

- In Chapter 3, we detailed our contribution on *"Joint Unicast Routing and Wireless Resource Allocation in Multihop LTE-D2D Communications"*, which put forward our ILP formulation of the routing and resource allocation problem while assuming only unicast traffic. Performance evaluation demonstrated that our proposal achieved good performances in terms of reliability, offloading ration, and latency in comparison to other basic single-sided optimal schemes.

- In Chapter 4, we gave an insight into our contribution on *"A Scalable Joint Routing and Resource Allocation Scheme: D2D-based Unicast and Multicast Data Offloading"*. As a first step, we put forward our ILP formulation to solve the underlying problem for a unified traffic model for both unicast and multicast traffic. To address the non-scalability of the initial formulation, we proposed a novel path-based ILP model in which a routing tree is expressed in terms of its path components. Next, for reason of speed, we proposed a sub-optimal solution method, based on the Column-Generation framework. In this formulation, we used a pricing problem specially modified to be more tractable to be solved by the fast Bellman-Ford algorithm. Performance evaluation revealed that our novel proposal achieved excellent performances in terms of reliability, latency, and scalability.

- In Chapter 5, we described our contribution *"D2D-Based Cellular Traffic Offloading: An Energy-Aware Scalable Heuristic Scheme"*. The energy-awareness, in this contribution, is addressed in two steps. First, we presented an optimal ILP-based approach that did not scale well with high-density topologies. Second, we presented a novel heuristic method composed of a parametric three-stage algorithm. Performance evaluation

shows that the presented heuristic outperformed the original one in terms of speed. As a result of massive speedups, up to six orders of magnitude, the heuristic scaled very well in denser topologies at the expense of performance gaps.

6.2 Future Work and Perspectives

In the short-term perspective, we plan to continue the work of this thesis along two principal axes. **Firstly**, we aim at studying further the optimality-speed compromise that we have confronted in our proposals. Indeed, enhancing the converging time is crucial to improve the scalability of the proposed offloading schemes. To do this, we must also consider decreasing the performance gaps relative to the optimal methods. **Secondly**, we intend to improve the evaluation methodology of thesis proposals by executing experimental testbeds. To this end, we consider using the **Open5G Lab** and **FlexRAN** platforms from the **Mosaic5G** open source ecosystem[1].

In the mid-term perspective, we outline more challenging topics related to this thesis. **First**, we propose to study the mobility issue in D2D relaying network. Including mobility in D2D multihop systems requires advanced study and analysis of the queues and their stability since D2D links can appear and disappear. This implies that the relaying mechanism becomes more opportunistic. **Second**, we propose to evaluate the security aspect in designing D2D relaying systems as in our offloading proposals. The existing end-to-end security scheme should be assessed within the framework of the current LTE-D2D standard. **Third**, an important issue to deal with is to convince and motivate (i.e., incentivize) the UEs to act as relays in such systems, which represents a

[1] See "http://mosaic-5g.io/"

significant business challenge for operators.

6.3 Publications

- **Journals**

  1. Safwan Alwan, Ilhem Fajjari, Nadjib Aitsaadi, Mejdi Kaddour, *"A Scalable Scheme for Joint Routing and Resource Allocation in LTE-D2D Based Offloading"*, IEEE Transactions on Network and Service Management, **under review**.
  2. Safwan Alwan, Ilhem Fajjari, Nadjib Aitsaadi, Paul Rubin, *"D2D-Based Cellular Traffic Offloading: An Energy-Aware Scalable Heuristic Scheme"*, IEEE Transactions on Networking, **under review**.

- **Conference Papers**

  1. Safwan Alwan, Ilhem Fajjari, Nadjib Aitsaadi, *"A Scalable Joint Routing and OFDMA Resource Allocation in LTE-D2D Networks"*, The 2019 IEEE Wireless Communications and Networking Conference (WCNC 2019), 15-18 April 2019, Marrakesh, Morocco.
  2. Safwan Alwan, Ilhem Fajjari, Nadjib Aitsaadi, *"Joint Routing and Wireless Resource Allocation in Multihop LTE-D2D Communications"*, The 43rd IEEE Conference on Local Computer Networks (LCN 2018), 1-4 October 2018, Chicago, USA.
  3. Safwan Alwan, Ilhem Fajjari, Nadjib Aitsaadi, *"D2D Multihop Energy-Efficient Routing and OFDMA Resource Allocation in 5G Networks"*, The IFIP Networking 2018 Conference (NETWORKING 2018), 14-16 May 2018, Zurich, Switzerland.

4. Safwan Alwan, Ilhem Fajjari, Nadjib Aitsaadi, *"Joint Multicast Routing and OFDM Resource Allocation in LTE-D2D 5G Cellular Network"*, The 2018 IEEE/IFIP Network Operations and Management Symposium (NOMS 2018), 23-27 April 2018, Taipei, Taiwan.

Bibliography

[1] V. Cisco, "Cisco visual networking index: Forecast and trends, 2017–2022," *White Paper*, vol. 1, 2018.

[2] E. Dahlman, S. Parkvall, and J. Skold, *5G NR: The next generation wireless access technology*. Academic Press, 2018.

[3] T. E. Bogale and L. B. Le, "Massive MIMO and mmwave for 5G wireless hetnet: Potential benefits and challenges," *IEEE Vehicular Technology Magazine*, vol. 11, no. 1, pp. 64–75, 2016.

[4] Y. Niu, Y. Li, D. Jin, L. Su, and A. V. Vasilakos, "A survey of millimeter wave communications (mmwave) for 5G: opportunities and challenges," *Wireless networks*, vol. 21, no. 8, pp. 2657–2676, 2015.

[5] P. K. Agyapong, M. Iwamura, D. Staehle, W. Kiess, and A. Benjebbour, "Design considerations for a 5G network architecture." *IEEE Communications Magazine*, vol. 52, no. 11, pp. 65–75, 2014.

[6] C. Westphal, "Challenges in networking to support augmented reality and virtual reality," *IEEE ICNC*, 2017.

[7] R. I. Ansari, C. Chrysostomou, S. A. Hassan, M. Guizani, S. Mumtaz, J. Rodriguez, and J. J. Rodrigues, "5g d2d networks: Techniques, challenges, and future prospects," *IEEE Systems Journal*, vol. 12, no. 4, pp. 3970–3984, 2017.

[8] 3GPP, "5G; service requirements for next generation new services and markets," 3rd Generation Partnership Project (3GPP), TS 22.261, Dec. 2018. [Online]. Available: http://www.3gpp.org/DynaReport/22261.htm

[9] H. Zhang, N. Liu, X. Chu, K. Long, A.-H. Aghvami, and V. C. Leung, "Network slicing based 5g and future mobile networks: mobility, resource management, and challenges," *IEEE Communications Magazine*, vol. 55, no. 8, pp. 138–145, 2017.

[10] X. Foukas, G. Patounas, A. Elmokashfi, and M. K. Marina, "Network slicing in 5G: Survey and challenges," *IEEE Communications Magazine*, vol. 55, no. 5, pp. 94–100, 2017.

[11] J. Ordonez-Lucena, P. Ameigeiras, D. Lopez, J. J. Ramos-Munoz, J. Lorca, and J. Folgueira, "Network slicing for 5g with sdn/nfv: Concepts, architectures, and challenges," *IEEE Communications Magazine*, vol. 55, no. 5, pp. 80–87, 2017.

[12] S. Li, L. Da Xu, and S. Zhao, "5G internet of things: A survey," *Journal of Industrial Information Integration*, vol. 10, pp. 1–9, 2018.

[13] S. Chen, J. Hu, Y. Shi, Y. Peng, J. Fang, R. Zhao, and L. Zhao, "Vehicle-to-everything (V2X) services supported by LTE-based systems and 5G," *IEEE Communications Standards Magazine*, 2017.

[14] R. Molina-Masegosa and J. Gozalvez, "LTE-V for sidelink 5G V2X vehicular communications: A new 5G technology for short-range vehicle-

to-everything communications," *IEEE Vehicular Technology Magazine*, vol. 12, no. 4, pp. 30–39, 2017.

[15] R.-A. Pitaval, O. Tirkkonen, R. Wichman, K. Pajukoski, E. Lahetkangas, and E. Tiirola, "Full-duplex self-backhauling for small-cell 5G networks," *IEEE Wireless Communications*, vol. 22, no. 5, pp. 83–89, 2015.

[16] Y.-D. Lin and Y.-C. Hsu, "Multihop cellular: A new architecture for wireless communications," in *Proceedings IEEE INFOCOM 2000. Conference on Computer Communications. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies (Cat. No. 00CH37064)*, vol. 3. IEEE, 2000, pp. 1273–1282.

[17] A. Asadi, Q. Wang, and V. Mancuso, "A Survey on Device-to-Device Communication in Cellular Networks," *IEEE Communications Surveys Tutorials*, 2014.

[18] K. M. G. Chavez, L. Goratti, T. Rasheed, D. B. Oljira, R. Fedrizzi, and R. Riggio, "The evolutionary role of communication technologies in public safety networks," in *Wireless Public Safety Networks 1*. Elsevier, 2015, pp. 21–48.

[19] T. Doumi, M. F. Dolan, S. Tatesh, A. Casati, G. Tsirtsis, K. Anchan, and D. Flore, "LTE for public safety networks," *IEEE Communications Magazine*, vol. 51, no. 2, pp. 106–112, February 2013.

[20] J. Liu, N. Kato, J. Ma, and N. Kadowaki, "Device-to-device communication in LTE-advanced networks: A survey," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 4, pp. 1923–1940, 2014.

[21] F. Rebecchi, M. D. De Amorim, V. Conan, A. Passarella, R. Bruno, and M. Conti, "Data offloading techniques in cellular networks: A survey," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 2, pp. 580–603, 2014.

[22] 3GPP, "Feasibility study for Proximity Services (ProSe)," 3rd Generation Partnership Project (3GPP), TR 22.803, 2013, v12.2.0.

[23] A. Aijaz, H. Aghvami, and M. Amani, "A survey on mobile data offloading: technical and business perspectives," *IEEE Wireless Communications*, vol. 20, no. 2, pp. 104–112, 2013.

[24] A. Pyattaev, K. Johnsson, S. Andreev, and Y. Koucheryavy, "Proximity-based data offloading via network assisted device-to-device communications," in *2013 IEEE 77th Vehicular Technology Conference (VTC Spring)*. IEEE, 2013, pp. 1–5.

[25] ——, "3GPP LTE traffic offloading onto WiFi direct," in *2013 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*. IEEE, 2013, pp. 135–140.

[26] G. Steri, G. Baldini, I. N. Fovino, R. Neisse, and L. Goratti, "A novel multihop secure lte-d2d communication protocol for iot scenarios," in *2016 23rd International Conference on Telecommunications (ICT)*, May 2016, pp. 1–6.

[27] O. Bello and S. Zeadally, "Intelligent device-to-device communication in the internet of things," *IEEE Systems Journal*, 2016.

[28] 3GPP, "Proximity-based services (ProSe); Stage 2," 3rd Generation Partnership Project (3GPP), TS 23.303, Dec. 2016. [Online]. Available: http://www.3gpp.org/DynaReport/23303.htm

[29] ——, "Study on lte device to device proximity services; radio aspects," 3rd Generation Partnership Project (3GPP), TR 36.213, Mar. 2014, v12.0.1. [Online]. Available: http://www.3gpp.org/DynaReport/36843.htm

[30] P. Gandotra, R. K. Jha, and S. Jain, "A survey on device-to-device (D2D) communication: Architecture and security issues," *Journal of Network and Computer Applications*, vol. 78, pp. 9–29, 2017.

[31] D. Griffith, A. B. Mosbah, and R. Rouil, "Group discovery time in device-to-device (D2D) proximity services (ProSe) networks," in *IEEE INFOCOM 2017-IEEE Conference on Computer Communications*. IEEE, 2017, pp. 1–9.

[32] A. Prasad, A. Kunz, G. Velev, K. Samdanis, and J. Song, "Energy-efficient D2D discovery for proximity services in 3GPP LTE-advanced networks: Prose discovery mechanisms," *IEEE vehicular technology magazine*, vol. 9, no. 4, pp. 40–50, 2014.

[33] S. Hakola, T. Chen, J. Lehtomaki, and T. Koskela, "Device-to-device (D2D) communication in cellular network-performance analysis of optimum and practical communication mode selection," in *2010 IEEE wireless communication and networking conference*. IEEE, 2010, pp. 1–6.

[34] R. Ma, N. Xia, H.-H. Chen, C.-Y. Chiu, and C.-S. Yang, "Mode selection, radio resource allocation, and power coordination in D2D communications," *IEEE Wireless Communications*, vol. 24, no. 3, pp. 112–121, 2017.

[35] M. Jung, K. Hwang, and S. Choi, "Joint mode selection and power allocation scheme for power-efficient device-to-device (D2D) communication," in *2012 IEEE 75th vehicular technology conference (VTC Spring)*. IEEE, 2012, pp. 1–5.

[36] S. Ali and A. Ahmad, "Resource allocation, interference management, and mode selection in device-to-device communication: A survey," *Transactions on Emerging Telecommunications Technologies*, vol. 28, no. 7, p. e3148, 2017.

[37] F. S. Shaikh and R. Wismüller, "Routing in multi-hop cellular device-to-device (D2D) networks: A survey," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 4, pp. 2622–2657, 2018.

[38] C. She, C. Yang, and T. Q. Quek, "Cross-layer optimization for ultra-reliable and low-latency radio access networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 1, pp. 127–141, 2017.

[39] S. Alwan, I. Fajjari, and N. Aitsaadi, "Joint routing and wireless resource allocation in multihop LTE-D2D communications," in *2018 IEEE 43rd Conference on Local Computer Networks (LCN)*. IEEE, 2018, pp. 167–174.

[40] ——, "Joint multicast routing and OFDM resource allocation in LTE-D2D 5G cellular network," in *IEEE/IFIP Network Operations and Management Symposium (NOMS)*, 2018.

[41] S. Alwan, I. Fajjari, and N. Aitsaadi, "D2D multihop energy-efficient routing and OFDMA resource allocation in 5G networks," in *2018 IFIP Networking Conference (IFIP Networking) and Workshops*, May 2018, pp. 1–9.

[42] A. Papadogiannis, E. Hardouin, A. Saadani, D. Gesbert, and P. Layec, "A novel framework for the utilisation of dynamic relays in cellular networks," in *2008 42nd Asilomar Conference on Signals, Systems and Computers*. IEEE, 2008, pp. 975–979.

[43] J. Li, W. Xia, S. Xing, and L. Shen, "Transmission scheduling and congestion control for multi-hop D2D underlaying cellular networks," in *2014*

IEEE 25th Annual International Symposium on Personal, Indoor, and Mobile Radio Communication (PIMRC), Sept 2014, pp. 996–1000.

[44] L. Babun, A. İ. Yürekli, and İ. Güvenç, "Multi-hop and D2D communications for extending coverage in public safety scenarios," in *2015 IEEE 40th Local Computer Networks Conference Workshops (LCN Workshops)*, Oct 2015, pp. 912–919.

[45] V. Bhardwaj and C. R. Murthy, "On optimal routing and power allocation for D2D communications," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 3063–3067.

[46] W. Cao, G. Feng, S. Qin, and Z. Liang, "D2D communication assisted traffic offloading for massive connections in HetNets," in *2016 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2016, pp. 1–6.

[47] C. Liu, C. He, and W. Meng, "A tractable multi-rats offloading scheme on d2d communications," *IEEE Access*, vol. 5, pp. 20 841–20 851, 2017.

[48] X. Zhang and Q. Zhu, "Statistical qos provisioning over d2d-offloading based 5g multimedia big-data mobile wireless networks," in *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, April 2018, pp. 742–747.

[49] G. Zhao, S. Chen, L. Qi, L. Zhao, and L. Hanzo, "Mobile-traffic-aware offloading for energy- and spectral-efficient large-scale d2d-enabled cellular networks," *IEEE Transactions on Wireless Communications*, vol. 18, no. 6, pp. 3251–3264, June 2019.

[50] R. O. Afolabi, A. Dadlani, and K. Kim, "Multicast Scheduling and Resource Allocation Algorithms for OFDMA-Based Systems: A Survey," *IEEE Communications Surveys & Tutorials*, 2013.

[51] F. Rebecchi, L. Valerio, R. Bruno, V. Conan, M. D. de Amorim, and A. Passarella, "A joint multicast/D2D learning-based approach to lte traffic offloading," *Computer Communications*, vol. 72, pp. 26–37, 2015.

[52] Y. Liu, A. M. A. E. Bashar, Fan Li, Y. Wang, and Kun Liu, "Multi-copy data dissemination with probabilistic delay constraint in mobile opportunistic device-to-device networks," in *IEEE International Symposium on A World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, 2016.

[53] Y. Xu and P. Wu, "Device-to-Device Multicast Content Delivery in Cellular Networks," in *EAI International Conference on Mobile Multimedia Communications*, ICST, Brussels, Belgium, Belgium, 2016.

[54] Z. Xia, J. Yan, and Y. Liu, "Energy efficiency in multicast multihop D2D networks," in *IEEE/CIC International Conference on Communications in China*, 2016.

[55] G. Rigazzi, F. Chiti, R. Fantacci, and C. Carlini, "Multi-hop D2D networking and resource management scheme for M2M communications over LTE-A systems," in *2014 International Wireless Communications and Mobile Computing Conference (IWCMC)*, Aug 2014, pp. 973–978.

[56] T. Ta, J. S. Baras, and C. Zhu, "Improving smartphone battery life utilizing device-to-device cooperative relays underlaying LTE networks," in *2014 IEEE International Conference on Communications (ICC)*, June 2014, pp. 5263–5268.

[57] B. Liu, Y. Cao, W. Wang, and T. Jiang, "Energy budget aware device-to-device cooperation for mobile videos," in *2015 IEEE Global Communications Conference (GLOBECOM)*, Dec 2015, pp. 1–7.

[58] A. Laha, X. Cao, W. Shen, X. Tian, and Y. Cheng, "An energy efficient routing protocol for device-to-device based multihop smartphone networks," in *2015 IEEE International Conference on Communications (ICC)*, June 2015, pp. 5448–5453.

[59] Z. Jingyi, L. Xi, and X. Quansheng, "Multi-hop routing for energy-efficiency enhancement in relay-assisted device-to-device communication," *The Journal of China Universities of Posts and Telecommunications*, vol. 22, no. 2, pp. 1–51, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S100588851560632X

[60] Cisco Visual Networking, "Cisco Global Cloud Index: Forecast and Methodology, 2015-2020," *White paper*, 2016.

[61] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE communications magazine*, vol. 52, no. 2, pp. 186–195, 2014.

[62] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, "Millimeter wave mobile communications for 5G cellular: It will work!" *IEEE access*, vol. 1, pp. 335–349, 2013.

[63] R. Wang, H. Hu, and X. Yang, "Potentials and challenges of C-RAN supporting multi-rats toward 5g mobile networks," *IEEE Access*, vol. 2, pp. 1187–1195, 2014.

[64] O. Galinina, A. Pyattaev, S. Andreev, M. Dohler, and Y. Koucheryavy, "5G multi-RAT LTE-WiFi ultra-dense small cells: Performance dynamics, architecture, and trends," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 6, pp. 1224–1240, 2015.

[65] V. Jungnickel, K. Manolakis, W. Zirwas, B. Panzner, V. Braun, M. Lossow, M. Sternad, R. Apelfrojd, and T. Svensson, "The role of small cells, coordinated multipoint, and massive MIMO in 5G," *IEEE communications magazine*, vol. 52, no. 5, pp. 44–51, 2014.

[66] P. Gandotra and R. K. Jha, "Device-to-Device Communication in Cellular Networks: A Survey," *Journal of Network and Computer Applications*, vol. 71, pp. 99–117, 2016.

[67] J. E. Mitchell, "Integer Programming: Branch and Cut Algorithms." in *Encyclopedia of Optimization*, C. A. Floudas and P. M. Pardalos, Eds. Springer, 2009.

[68] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures (Release 14)," 3rd Generation Partnership Project (3GPP), TS 36.213, Sep. 2017, v14.4.0. [Online]. Available: http://www.3gpp.org/DynaReport/36213.htm

[69] G. F. Riley and T. R. Henderson, "The NS-3 Network Simulator Modeling and Tools for Network Simulation," in *Modeling and Tools for Network Simulation*, K. Wehrle, M. Güneş, and J. Gross, Eds. Springer Berlin Heidelberg, 2010, ch. 2.

[70] N. Baldo, M. Miozzo, M. Requena-Esteso, and J. Nin-Guerrero, "An open source product-oriented LTE network simulator based on NS-3," in *Pro-*

ceedings of the 14th ACM international conference on Modeling, analysis and simulation of wireless and mobile systems. ACM, 2011, pp. 293–298.

[71] Y. d. J. Bultitude and T. Rautiainen, "IST-4-027756 WINNER II D1. 1.2 V1. 2 WINNER II Channel Models," 2007.

[72] H. Tullberg, P. Popovski, Z. Li, M. A. Uusitalo, A. Hoglund, O. Bulakci, M. Fallgren, and J. F. Monserrat, "The metis 5g system concept: Meeting the 5g requirements," *IEEE Communications magazine*, vol. 54, no. 12, pp. 132–139, 2016.

[73] C. H. Papadimitriou and K. Steiglitz, *Combinatorial optimization*. Prentice Hall Englewood Cliffs, 1982, vol. 24.

[74] C. Barnhart, E. L. Johnson, G. L. Nemhauser, M. W. Savelsbergh, and P. H. Vance, "Branch-and-price: Column generation for solving huge integer programs," *Operations research*, vol. 46, no. 3, 1998.

[75] J. Desrosiers and M. Lübbecke, *A Primer in Column Generation*. Springer Science & Business Media, 03 2006, pp. 1–32.

[76] C. Hwang, S. Paidy, A. Masud, and K. Yoon, *Multiple Objective Decision Making — Methods and Applications: A State-of-the-Art Survey*, ser. Lecture Notes in Economics and Mathematical Systems. Springer Berlin Heidelberg, 2012.

[77] A. Schrijver, *Combinatorial optimization: polyhedra and efficiency*. Springer Science & Business Media, 2003, vol. 24.

[78] M. Hoeyhtyae, A. Maemmelae, U. Celentano, and J. Roening, "Power-efficiency in social-aware d2d communications," in *European Wireless 2016; 22th European Wireless Conference*, May 2016, pp. 1–6.

[79] M. Lauridsen, L. Noël, T. B. Sørensen, and P. Mogensen, "An empirical LTE smartphone power model with a view to energy efficiency evolution," *Intel Technology Journal*, vol. 18, no. 1, pp. 172–193, 2014.