UNIVERSIDAD CARLOS III DE MADRID

TESIS DOCTORAL

ANALYTICAL CHARACTERIZATION OF IN-BAND AND OUT-BAND D2D COMMUNICATIONS FOR NETWORK ACCESS

Autor: Christian Vitale, IMDEA Networks Institute, University Carlos III of Madrid
Director: Vincenzo Mancuso, IMDEA Networks Institute

DEPARTAMENTO DE INGENIERÍA TELEMÁTICA

Leganés (Madrid), Junio de 2017

UNIVERSIDAD CARLOS III DE MADRID

PH.D. THESIS

ANALYTICAL CHARACTERIZATION OF IN-BAND AND OUT-BAND D2D COMMUNICATIONS FOR NETWORK ACCESS

Author: Christian Vitale, IMDEA Networks Institute, University Carlos III of Madrid
Director: Vincenzo Mancuso, IMDEA Networks Institute

DEPARTMENT OF TELEMATIC ENGINEERING

Leganés (Madrid), June 2017

Analytical characterization of in-band and out-band D2D communications for network access

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Prepared by
Christian Vitale, IMDEA Networks Institute, University Carlos III of Madrid

Under the advice of
Vincenzo Mancuso, IMDEA Networks Institute
Departamento de Ingeniería Telemática, Universidad Carlos III de Madrid

---

Date: Junio, 2017

Web/contact: christian.vitale@imdea.org

This work has been supported by IMDEA Networks Institute.

TESIS DOCTORAL

ANALYTICAL CHARACTERIZATION OF IN-BAND AND OUT-BAND D2D COMMUNICATIONS FOR NETWORK ACCESS

Autor: Christian Vitale, IMDEA Networks Institute, University Carlos III of Madrid
Director: Vincenzo Mancuso, IMDEA Networks Institute

Firma del tribunal calificador:

Presidente: Prof. Marco Ajmone Marsan

Vocal: Prof. Omer Gurewitz

Secretario: Prof. Miquel Payaró

Calificación:

Leganés, de de

Acknowledgements

The doctorate is a trip, they say. Well, mine it has been an adventurous and crowded one. Many people participated to this effort. My supervisors, my family and friends they were all part of this long and exciting journey.

Vorrei cominciare ringraziando mia madre e mio padre. Non ci sono parole per descrivere quello che hanno fatto per me, quello che mi hanno insegnato. Sono la persona che sono grazie a loro. Spero di aver preso almeno un pó della tenacia, della caparbietá e del forte senso del lavoro di mia madre, un pó della pazienza o della capacitá di sorprendersi di mio padre. Avervi lontano per cosí tanti anni non é stato facile. Vorrei ringraziare anche mio fratello, per il suo profondo affetto e per la sua costante ammirazione. È una fortuna essere il suo fratello maggiore.

Next, I'll like to thank my three supervisors Gianluca, Balaji and Vincenzo. Gianluca and Balaji for their help in my first steps during the doctorate, for having the patience and the courage to listen to my first “presentations” or “ideas”. I know it wasn't easy. I would like to say a big thank you to Vincenzo, which gave me the possibility to be his student after that Gianluca and Balaji left IMDEA Networks. His teachings and his guidance helped me to gain a scientific thinking. Without him, this doctorate wouldn't be possible. I would like to thank also Jim, which I had the honor to work with during my internship. I never found in my career someone so successful and at the same time humble as him.

Moreover, I would like to thank my colleagues at IMDEA Networks for making this doctorate an unforgettable experience. Many of them are now my friends, part of my life, and I can't thank them enough for the support and for the shared memories. I would like to thank my desk mates Maurizio, Roberto, Camillo, Foivos and Philippe. They always had 5 minutes for me when I needed to talk and detach from work. I don't know how they could stand me all those years singing and mumbling while I was trying to focus. I wouldn't do the same for you guys. I would like to thank the guys of the IMDEA Networks futsal team. I hope that together we started a new tradition, something that it will continue with the years. After all we won the trophy for the best international team, right? (Ok, maybe we were the only international team) I would like to thank Nicholas, my Greek “kumparos” Angelos and their families for the invaluable company in all the Spanish bars we dared to enter during the long, and warm, summer nights. I would like to thank Dario, even if he cheers for Inter. The continuous futile arguments we have are the spice of life. I would like to thank Miguel for his bravery. Following us to salsa lessons for sure is the sign of

a true and, I hope, everlasting friendship. I would like to thank Pablo, Vincenzo and Arash for their collaboration, their efforts and for the sometimes never-ending Skype calls. I would also like to thank my friends back in Italy for their support. Marco, my Italian “compare”, Marco and Christian. Our life's path have brought us far, but every time we meet is like we never part each other. I would like to thank Gaia, Alessia, Aymen, Roderick, Nicola and all the rest which played an important role in this adventurous trip.

Finally, I would like to thank Ev, my inspiration. Cheerful, lively, sometimes on the verge of craziness, she arrived in my life like a typhoon but, somehow, she saved me. She has been my relief in this stressful time and I thank her for her patience and comprehension. Without her continuous support and her steady belief in me, this doctorate wouldn't have been possible.

Abstract

Cooperative short-range communication schemes provide powerful tools to solve interference and resource shortage problems in wireless access networks. With such schemes, a mobile node with excellent cellular connectivity can momentarily accept to relay traffic for its neighbors experiencing poor radio conditions and use Device-to-Device (D2D) communications to accomplish the task. This thesis provides a novel and comprehensive analytical framework that allows evaluating the effects of D2D communications in access networks in terms of spectrum and energy efficiency. The analysis covers the cases in which D2D communications use the same bandwidth of legacy cellular users (in-band D2D) or a different one (out-band D2D) and leverages on the characterization of underlying queueing systems and protocols to capture the complex intertwining of short-range and legacy WiFi and cellular communications.

The analysis also unveils how D2D affects the use and scope of other optimization techniques used for, e.g., interference coordination and fairness in resource distribution. Indeed, characterizing the performance of D2D-enabled wireless access networks plays an essential role in the optimization of system operation and, as a consequence, permits to assess the general applicability of D2D solutions. With such characterization, we were able to design several mechanisms that improve system capabilities. Specifically, we propose bandwidth resource management techniques for controlling interference when cellular users and D2D pairs share the same spectrum, we design advanced and energy-aware access selection mechanisms, we show how to adopt D2D communications in conjunction with interference coordination schemes to achieve high and fair throughputs, and we discuss on end-to-end fairness—beyond the use of access network resources—when D2D communications is adopted in C-RAN. The results reported in this thesis show that identifying performance bottlenecks is key to properly control network operation, and, interestingly, bottlenecks may not be represented just by wireless resources when end-to-end fairness is of concern.

Table of Contents

List of Tables

List of Figures

List of Acronyms

5G next generation of cellular networks

ABSF Almost-Blank SubFrame

ABSF-PF Almost-Blank SubFrame - Proportional Fairness

AP Access Point

Branch and Bound

BMF Bottleneck Max Fairness

BASICS BAse Station Inter-Cell Scheduling

BS Base Station

BSB Base Stations Blanking

C-RAN Centralized-Radio Access Network

CDF Cumulative Distribution Function

CPS Coupled Processor System

CRS Cell-specific Reference Symbols

CSI Channel State Information

CSMA-CA Carrier Sense Multiple Access - Collision Avoidance

D2D Device-to-Device

DRONEE Dual Radio Opportunistic Networking for Energy Efficiency

DRF Dominant Resource Fairness

ECN Explicit Congestion Notification

eD2D enhanced D2D

eNB evolved Node B

FFR Full Frequency Re-Use

FIFO First-In First-Out

GAP Generalized Assignment Problem

GO Group Owner

GM Group Member

GPS Generalized Processor Sharing

ICIC Inter Cell Interference Coordination

ICI Inter Cell Interference

IP Internet Protocol

ISD Inter-site Distance

JFI Jain's Fairness Index

LTE Long Term Evolution

LTE-A Long Term Evolution-Advanced

MBH Marginal Benefit Heuristic

MCS Modulation and Coding Scheme

MIMO Multiple-Input Multiple-Output Transmissions

mm-wave millimiter-wave

MN Mobile Node

NC Network Calculus

OFDM Orthogonal Frequency-Division Multiplexing

OFDMA Orthogonal Frequency-Division Multiple Access

PoA Point of Access

PDF Probability Density Function

PF Proportional Fair

ProSe Proximity-based Services

QoS Quality of Service

RAN Radio Access Network

RAT Radio Access Technology

RB Resource Block

RPGM Reference Point Group Mobility

RR Round Robin

RWP Random Way Point

SC-FDMA Single Carrier-FDMA

SDN Software Defined Network

SFQ Start-Time Fair Queuing

SINR Signal-to-Interference-plus-Noise Ratio

SIR Signal-to-Interference Ratio

SNR Signal-to-Noise Ratio

TCP Transmission Control Protocol

TOMRAN Two-level Opportunistic Spectrum Management for 5G Radio Access Networks

UE User Equipment

VFS Vertical First Scheduler

WiGig Wireless Gigabit

WLAN Wireless Local Area Network

WRR Weighted Round Robin

Chapter 1

Introduction

The promise of the next generation of cellular networks (5G) is to offer an unprecedented experience to end users. This includes extremely high data rates, very low latency and massive connectivity [1]. In order to achieve such competitive goals, the Radio Access Network (RAN) design has to change radically on the physical layer, on the protocols functioning and on the radio resource management. Furthermore, 5G RANs are envisioned to be used for many diverse applications. As a result, 5G RANs are going to support a wide range of spectrum bands and communication technologies, from to and beyond, with heterogeneous data channel bandwidths and propagation conditions. For example, Long Term Evolution-Advanced (LTE-A) technology will be also granted compatibility in novel 5G radio, together with WiFi evolutions. New spectrum will be available for 5G, including in the millimiter-waves (mm-waves) band. The joint usage of high and low spectrum frequencies combines the benefit of beamforming capabilities at higher bandwidths with the good coverage capabilities of low frequencies.

The high data-rate requirement for 5G calls for lower frequency re-use factor, higher spectral efficiency, and significantly denser deployments of base stations [2]. Therefore, intelligent small cells with the ability to make local decisions on the usage of resources or spectrum bands will complement legacy macro cells coverage. However, this trend towards network densification leads to challenging issues due to severe Inter Cell Interference (ICI) effects.

Several techniques have been proposed to cope with interference or low spectral efficiency in the next generation of cellular networks. 5G networks will leverage interference cancellation, beamforming and many other optimization mechanisms [3]. Each of the proposed techniques has its own peculiar advantage and disadvantage. When interference cancellation is in place, for example, a base station regenerates the useful transmissions received estimating and subtracting the interfering signals. Most of the times, low latency intra-base station communications are needed in order to correctly estimate interfering signals and decode useful information. While uplink application of interference cancellation techniques is a mature technology, even more if decoding operations are not performed at the base stations but performed in a central data-center, it is much more complicated to envision a user device estimating interfering signals without the

knowledge of the scheduling decisions of neighboring base stations. In beamforming, instead, the complexity of identifying the correct location of user terminals grows exponentially with the number of users to serve or with the number of base stations to simultaneously coordinate. Extreme semplifications have to be therefore introduced in the ultra-dense scenarios envisioned for 5G networks.

In this thesis we focus on one of the possible solutions to ICI, i.e. Device-to-Device (D2D) communications. The increasing penetration of interconnected devices (smartphones, sensors, home appliances, etc.) is giving progressively rise to a wealth of new services and possibilities. Millions of devices have nowadays cellular connectivity and multi-technology capabilities. Therefore, D2D can be used not only as a new pervasive communication technology but also as a new way of improving end user performance. In traditional D2D, two terminals in proximity with information to exchange transmit directly to each other without traversing the Base Station (BS) or the core network. In the standards, D2D communications are known as Proximity-based Services (ProSe) [4, 5] and their main goal is to offload the cellular infrastructure. Nevertheless, D2D communications have been also proposed as a relay technique to counteract ICI. Indeed, excellent cellular conditions of some users might be exploited to relay traffic for other neighboring nodes with disruptive channel conditions. This thesis presents a first analytical framework to analytically assess the possible gain that the introduction of D2D communications may bring to cellular networks both when used as a direct link among terminals and as a relay methodology. The evaluation performed is in terms of energy efficiency, fairness and aggregate throughput.

Different types of D2D schemes have been proposed up to now. *In-band* schemes either allow D2D transmissions to occur over dedicated cellular resources (*in-band overlay* schemes [6]) or over the same resources used by legacy cellular users (*in-band underlay* schemes [7]). The analysis of the performance of in-band schemes is challenging, due to the complex interaction of D2D and legacy cellular transmissions. In in-band underlay D2D, D2D and cellular users all share the same resources and their performance is strongly correlated through interference, through traffic patterns, and through the specific scheduling algorithm adopted. In in-band overlay D2D, instead, the challenge is in dynamically reserving to D2D transmissions the correct portion of the available resources. In order to avoid many of the above mentioned problems, *out-band* schemes for D2D communications have been also proposed. In out-band D2D, transmissions among terminals happen on a different bandwidth than cellular transmissions. Most commonly, out-band D2D communications happen on the bandwidths (at 2.4/5 GHz. or at 60 GHz.).

In the first part of this thesis, we start by characterizing the performance of several D2D communication schemes. In particular, among the possible approaches, we focus on in-band underlay and out-band D2D communications. Both transmission schemes are indeed very promising. In-band underlay D2D allows scheduling several transmissions at the same time on the same resources. Therefore, if done properly, it allows to largely improve spectrum efficiency thanks to the exploitation of spatial diversity. Cellular spectrum is quite scarce nowadays, and in-band underlay D2D would therefore allow increasing the overall aggregate throughput achieved within the

access network. Out-band D2D communications, instead, allows increasing the overall throughput using a different bandwidth than the cellular one. On one hand, using a different bandwidth grants avoiding the entangled interference management that in-band underlay D2D requires. On the other hand, out-band underlay D2D communications happen on a not-licensed bandwidth and it could be insufficient to deliver the Quality of Service (QoS) that a cellular user nowadays expects. In this thesis we show that there is not a clear winner among in-band underlay D2D communications and out-band D2D communications and cellular operators may choose among those two options depending on the particular situation at hand. Nevertheless, leveraging the presented analytical framework, we propose novel mechanisms which improve the throughput fairness and the energy efficiency both in out-band and in-band underlay D2D communication schemes.

After analysing the performance of D2D communications in RANs, both as a direct link among users and as a relay methodology, in the second part of this thesis we evaluate the effects of D2D when the cellular network is already implementing other Inter Cell Interference Coordination (ICIC) techniques. So far, the majority of research efforts have taken simplifying assumptions and tackled interference or low spectral efficiency in isolation. Unfortunately, most of the resulting techniques have conflicting or overlapping objectives, whose compound effects have been rarely evaluated. In this thesis instead, we propose an analytical framework which takes into account a RAN where D2D is deployed simultaneously with Almost-Blank SubFrame (ABSF). ABSF represents a 3GPP standard technique to partially mute some base stations when ICI is above a threshold. ABSF mechanism prevents the base station (e.g., the evolved Node B (eNB) of Long Term Evolution (LTE)/LTE-A networks) from transmitting in a specified set of subframes. ABSF reduces the overall sensed interference and allows for using more efficient Modulation and Coding Schemes (MCSs), achieving higher throughputs. Thanks to the analytical framework presented, we show how to jointly optimize D2D and ABSF in order to improve network aggregate throughput or per-user fairness.

Finally, we present a new methodology to enforce fairness in D2D-enabled Centralized-Radio Access Networks (C-RANs). C-RANs have been proposed as a possible solution to reduce the management complexity which is due to the densification of access networks. A C-RAN consists of multiple radio Points of Access (PoAs) performing only signal transmission/reception, a fronthaul/backhaul connection and a central data center. The data center manages a pool of virtual base stations, each representing one or more actual base stations, which perform the majority of the base-band processing. In case of D2D-enabled C-RANs, the data center is also responsible of the management of the D2D relay groups. As a result, in D2D-enabled C-RANs flows may have to traverse several resources before reaching destination. In the wired Internet the resources to traverse are links and all flows have the same requirement for each bit/s of rate. In C-RANs, though, the resources traversed by the flows are wireless links and computational resources and requirements may be heterogeneous. For example, the amount of spectrum consumed for each bit/s depends on the flow's radio conditions. In the same way, the amount of CPU/RAM required for encoding/decoding depends on the number of information bits transmitted per each

Orthogonal Frequency-Division Multiplexing (OFDM) symbol [8] and, somehow, on radio conditions. Resource sharing by flows with homogeneous requirements has been widely studied over many years. Our main objective is to derive the equivalent result in the case of heterogeneous resource requirements. Explicitly, we show that imposing local fairness at each resource, coupled with end-to-end flow control, results in a desirable generalization of max-min fairness called Bottleneck Max Fairness (BMF) [9]. A BMF allocation is such that resource sharing is Pareto efficient (i.e., all capacity is used if possible) and every flow receives the maximum allocation at some resource that is fully used. Furthermore, BMF allocations are very close to proportional fair resource allocations.

1.1. Main Contributions and Organization of the Thesis

In this section we highlight the main findings that can be found in this thesis and its organization.

Chapter 2 discusses the related work and introduces some basic concepts of Network Calculus (NC), which are then used in the following.

In the first part of this thesis we analyze the performance of a RAN when D2D communications are used as a direct transmission link among users or with relay purposes. We take into consideration both in-band underlay and out-band D2D communications, proposing the use of a common underlying analytical framework. In both cases indeed, the throughput that a given transmitter is able to deliver to the destination is univocally determined by the set of concurrent active transmitters. In in-band underlay D2D, the set of transmitters which have been scheduled on the same frequency and at the same time univocally determines the interference sensed at destination, and therefore the achievable throughput. The same holds for out-band D2D transmitters. As shown in [10], the probability of successful transmission, as well as the probability of colliding with other transmitters, is univocally determined by the set of devices with traffic to send. Therefore, in order to study the performances of the D2D systems, we consider a queuing system for which the service rate of each queue depends on the set of queues with ongoing job service. Such model is commonly referred to as Coupled Processor System (CPS) [11–13], and it arises also in several other contexts in which coupling derives from resource sharing.

Unfortunately, CPS performance analysis is as complex as wireless performance analysis. Up to now, the characterization of the delay experienced by the CPS's arrivals, as well the achievable service rates are known only for a 2-queue CPS. Chapter 3 presents an innovative method to study CPS performance, and subsequently, to study the performance of D2D communications. We propose a fully analytical approach to the performance analysis of a generic -queue CPS, which does not require simulations to be parametrized. The proposed method is independent of the nature and shape of coupling relations. Furthermore, our method can be applied to systems in which the traffic is described either stochastically or by means of deterministic arrival curves [14].

Given some prior knowledge on the characterization of traffic arrivals, Chapter 3 allow com-

puting a conservative estimation of the throughput achieved by D2D transmitters both in case of in-band underlay and out-band D2D. Specifically, Chapter 4 shows how to evaluate the achievable network aggregate throughput when a particular in-band underlay scheduling policy is used, i.e., FlashLinQ [15]. FlashLinQ is a state-of-the-art PHY-MAC architecture for D2D, promoted by Qualcomm, that allows the scheduling of different transmitters (D2D or cellular) in the same time and frequency resources, through an Orthogonal Frequency-Division Multiple Access (OFDMA)-like access selection mechanism. FlashLinQ is one of the most important candidates to be adopted by 5G as the scheduling protocol of in-band D2D communications. Chapter 5, instead, studies the performance of the Carrier Sense Multiple Access - Collision Avoidance (CSMA-CA) contention mechanism. bandwidth is indeed the most probable candidate to be used for out-band D2D communications to offload those transmissions among users which instead should traverse the whole access network. Many of the available results are typically valid under the assumption that the incoming traffic at the transmitters is stationary and Poissonian. This has been done despite the fact that traffic in real networks is well far from being Poissonian (see [16] and related literature). In particular, traffic from live audio/video streaming exhibits a periodic behavior which substantially departs from the Poisson model. Chapter 5 presents an analysis which needs little knowledge about the arrivals traffic statistics.

The knowledge of the set of achievable rates by the D2D users it is important not only because it allows understanding what are the actual performance achieved, but also because it allows to tune the operating point of the D2D transmitters. With simple rate limiting techniques it is possible to optimize over the set of achievable rates the throughput of the D2D transmitters in the system. In Chapter 4 and Chapter 5 we show a simple but effective methodology to achieve proportional fairness.

After evaluating, and tuning, the performance of in-band underlay and out-band D2D communications when used as direct transmission link among cellular users, in Chapter 6 we present an analytical framework to evaluate the performance achieved by a cellular operator which decides to use D2D communications as a relay methodology. Herein, we assume that Telco operators implement alongside the traditional cellular infrastructure on one hand -based hotspots [17] and, on the other hand, D2D-assisted cooperative relay transmissions [4]. We name such access networks, *mixed-access networks*. In this context, D2D-assisted cellular access networks might allow bundling user traffic in groups and implementing dynamic relay, without requiring additional infrastructure deployments. Cooperative relay improves the end user experience towards destinations with poor or no channel quality. Therefore *D2D can be seen as an alternative network access technology.* Nevertheless, including in the picture also hotspots not only allows us to capture the tendency of operators to place side by side cellular and connectivity, but it also allows us to evaluate if static hotspots may suffice to offload the cellular access network, making redundant the use of D2D cooperative relay. In Chapter 6, we derive a novel comprehensive model for throughput and terminal power consumption estimation in mixed-access networks which takes into account the intertwined nature of cellular and resource alloca-

tion when D2D comes into play. While acting occasionally as relay, mobile devices experience a higher power consumption [18, 19] since they have to receive and re-transmit the traffic of other User Equipments (UEs). Therefore, battery lifetime becomes of paramount importance in mixed-access networks. Building on our model, we formulate a new network-controlled access selection problem that aims at maximizing the energy efficiency achieved at the terminal side. As a major point, in the proposed formulation, not only uplink and downlink are both simultaneously considered, but they are also considered as jointly optimized in terms of cellular and resource allocation.

In the second part of this thesis, we analyze the performance of a cellular network which simultaneously deploys D2D communications and other different optimization techniques.

In Chapter 7 we unveil a first control mechanism which deals *jointly* with interference and spectrum efficiency. We call such a control mechanism Two-level Opportunistic Spectrum Management for 5G Radio Access Networks (TOMRAN). We focus on two particular techniques, namely an ICIC scheme based on the ABSF paradigm for interference management, and an out-band D2D relay strategy for clusters of collaborative neighbors to improve spectral efficiency. Compared to other interference management approaches, the ABSF concept is simple enough to be accepted for inclusion into technically entangled 3GPP specifications and at the same time it has found rather wide acceptance among standardization institutions [20]. Out-band D2D communications, instead, avoid unnecessary traffic in the cellular infrastructure and therefore represent the key-enabler to cope with low spectral efficiency in the future high-density network design. Specifically, TOMRAN jointly coordinates two state-of-the-art techniques to group users and to perform ABSF, a.k.a., Dual Radio Opportunistic Networking for Energy Efficiency (DRONEE) [21] and Base Stations Blanking (BSB) [22], respectively. DRONEE forms groups by means of a *Merge & Split* algorithm which targets the maximization of per-user throughput. BSB, instead, optimize the patterns of activity of each base station to guarantee a minimum SINR level to any transmission in the cellular network. In Chapter 7, we evaluate the performance of TOMRAN with an extensive simulation campaign. The results suggest that, even if a standard ABSF technique as BSB and out-band D2D have overlapping objectives, when used altogether, like in TOMRAN, they largely outperforms a mechanism which instead takes in consideration only one of the two.

Chapter 8 picks up on the the limitations of ICIC approaches like ABSF in heterogeneous scenarios and shows that D2D can be used to boost throughputs selectively. ABSF, indeed, is a PHY-layer technique and throughput fairness is naturally out of its design scope. However, ABSF can be seen as a tool for scheduling base station activities, hence can be used to enforce fairness among cells. Moreover, we adopt a technological solution ready to serve the next generation of D2D relay communications: mm-waves communications facilities targeted by the IEEE ad standard, commonly known as Wireless Gigabit (WiGig) [23]. WiGig achieves virtually unlimited speeds with respect to the cellular capacity, thus making enhanced relay groups a solution one step beyond classical D2D. The compound impact of ABSF and relay consists in regulating

the number of simultaneous cellular transmissions while at the same time reducing the presence of vulnerable users experiencing poor channel conditions. Hence, fairness and transmission efficiency can be jointly targeted by simultaneously controlling inter-cell activity (with ABSF) and intra-cell packet relay (with enhanced D2D (eD2D)) within a novel unified framework. To fully understand the potentials of ABSF and eD2D in combination, and to make the basis for advanced control architectures for 5G networks, we present a theoretical study on the limitations of ABSF and on the potentials of eD2D, derive stochastic conditions to show how ABSF can be used to steer user fairness, formulate novel and convex optimization problems to set stochastic ABSF activity patterns by leveraging the advantages of eD2D relay and show that the joint operation of stochastic ABSF and eD2D is practical and brings dramatic gains with respect to state-of-the-art solutions.

Finally, in Chapter 9, we go a step beyond the optimization of the last wireless hop and we consider a RAN where coding and decoding of received signals is done in a centralized datacenter, i.e., D2D-enabled C-RANs. Explicitly, we show that imposing local fairness at each resource a flow has to traverse in C-RANs, coupled with end-to-end flow control, results in a desirable generalization of max-min fairness called BMF [9]. Proving this result is hard due to the complex dynamics of per-flow resource queue backlogs. We apply a fluid model to prove convergence to BMF for flows with heterogeneous requirements in a network limited to two resources. It is considerably harder to account for heterogeneous requirements because the water filling characterization of max-min fairness used does not generalize. After the mathematical analysis enabled by the fluid model adopted, first we evaluate the approximation introduced to the theoretical objective of multi-resource sharing in a generic packet-based scenario. Our goal is indeed understanding how a realistic assumption on the arrival traffic deviates the real system performance from its ideal fluid behavior. We simulate a network where resources implement Start-Time Fair Queuing (SFQ) [24] and investigate the impact on convergence times of window size, the number of competing flows and their particular requirements. Finally we apply local fairness at each resource in D2D-assisted C-RAN and we evaluate the achieved performance against the proportional fairness algorithm.

Finally, we summarize the scientific output of this thesis and we briefly discuss future research directions in Chapter 10.

1.2. Publications

The work presented in this doctoral thesis has appeared in six conference papers [25–30] and one poster [31]. Additionally, one more conference paper is under review. In this section we briefly elaborate on these publications and how they form part of the doctoral thesis.

We have published our work related to the evaluation of the performance of in-band underlay and out-band D2D in different conference papers (listed below). Each of the paper characterizes the set of achievable throughputs by the D2D transmitters in one of the mentioned transmission

modes. Building on the proposed modeling, it was also illustrated a rate-limiting technique able to select the set of throughputs for the D2D transmitters that are sustainable by the system and that optimize the proportional fairness of the per-user throughputs.

- C. Vitale, V. Mancuso, and G. Rizzo, "Modelling D2D communications in cellular access networks via coupled processors" in 7th International Conference on Communication Systems and Networks, COMSNETS 2015, Bangalore, India, January 6-10, 2015, 2015, pp. 1 8.

- C. Vitale, G. Rizzo, and V. Mancuso, "A coupled processors model for 802.11 ad hoc networks under non-saturation" in 2015 IEEE International Conference on Communications, ICC 2015, London, United Kingdom, June 8 12, 2015, 2015, pp. 628 634.

Please note that our COMSNETS'15 publication was also awarded of the "Honorable Mention Award", i.e., the price for the third best paper of the conference. The organizers of the conference also invited us to present a poster at the same event. The modeling underlying the analysis of in-band underlay and out-band D2D was based on the CPS queueing system. The CPS is able to capture the complex coupling (due to interference or to the MAC of WiFi) among throughput performance of D2D transmitters. Our contribution to the modeling of CPS was presented in the following article.

- C. Vitale, G. Rizzo, B. Rengarajan, and V. Mancuso, "An analytical approach to performance analysis of coupled processor systems," in 27th International Teletraffic Congress, ITC 2015, Ghent, Belgium, September 8-10, 2015, 2015, pp. 89 97.

In this thesis we also analyzed the possible performance of D2D-assisted cellular networks. In such a system, D2D communications are used to relay the traffic of users experiencing poor channel quality in order to improve the spectral efficiency. In our contribution, we analytical evalute what it would be the achieved throughput and the experienced power consumption by users' terminals when WiFi-Direct D2D communications are among the viable access alternatives. Thanks to the analytical framework introduced, we propose an access selection mechanism that maximize the energy efficiency of users' terminals. The contribution is part of the following article.

- C. Vitale and V. Mancuso, "Energy efficiency in mixed access networks," in Proceedings of the 19th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems, MSWiM 2016, Malta, November 13-17, 2016, 2016, pp. 35 42.

In the second part of the thesis, instead, we focus on the performance of D2D communications when they are jointly deployed with different other optimization techniques. First, we focus on the possible use of D2D communciations with ICIC mechanisms, such as ABSF. In this thesis, we first shed some lights, through a simulation study, on the possible gains achieved when users group

via WiFi-Direct and standard ABSF is in place. Then, we developed an analytical framework to systematically coordinate ABSF operations when users can group via WiGig. The fundamental difference here is that the capacity of the groups is much larger than the possible capacity of the cellular network, and groups are never a performance bottleneck. The analytical framework shows the intrinsic limitations of ABSF (alone) to improve aggregate throughput (even in heterogeneous environments), and suggests that ABSF would be much more useful if used to infer fairness. The results contained in this thesis have been published in:

- C. Vitale, V. Sciancalepore, A. Asadi, and V. Mancuso, "Two-level opportunistic spectrum management for green 5G radio access networks," in 2015 IEEE Online Conference on Green Communications, OnlineGreenComm 2015, Piscataway, NJ, USA, November 10-12, 2015, 2015, pp. 78 83,

and they are under submission at GLOBECOM'17. The work has been also awarded of a "Best Paper Award" at OnlineGreenComm 2015.

Finally, we analyzed a possible technique which allows to infer multi-resource sharing. Multi-resource sharing is fundamental in D2D-enabled C-RANs. Here base stations only have transmission/reception capabilities, ad-hoc datacenters perform base-band processing and relay nodes act as a mediator among the cellular infrastructure and the final destination of a traffic flow. In multi-resource sharing, depending on the requirements of the flows, any of the resources may represent the performance bottleneck of traffic flows and it is of fundamental importance to infer end-to-end fairness. The related results have been published in the following article.

- T. Bonald, J. Roberts, and C. Vitale, "Convergence to multi-resource fairness under end to end window control," The 36th Annual IEEE International Conference on Computer Communications (INFOCOM), 2017.

Chapter 2

Related Work and Background

In this section we review the most relevant works related to this thesis. We start comparing the work presented in this thesis with different other works dealing with in-band and out-band Device-to-Device (D2D) communications, with Inter Cell Interference Coordination (ICIC) and multi-resource sharing. Finally, we present some definitions and basic results of Network Calculus (NC) used in our analysis (especially in Chapter 3)

2.1. Related Work

2.1.1. In-band D2D Performance Evaluation

To date, few works have tried to characterize the performance of in-band underlay D2D communication schemes, both when D2D is used as a direct link among users or as relay methodology. The analysis of the performance of in-band underlay schemes is challenging, due to the complex interaction among D2D transmissions and cellular system operations. As D2D and cellular users all share the same resources, indeed, their performance is strongly correlated through interference, through their traffic patterns, and through the specific scheduling algorithm adopted. The typical assumption is that the system is in saturation [32] or performance evaluation is performed via simulations [33] or that only a single D2D pair can be scheduled on the same time and frequence together with a cellular user [7].

This leads to pessimistic, overly conservative results, particularly in non-saturated settings, and in general it does not enable the characterization of the main performance trade-offs of such D2D systems, essential for the design of efficient scheduling and rate allocation algorithms. Given an in-band underlay D2D communication scheme, in this thesis we present two novel results:
a method that allows to know if the traffic arriving at the D2D transmitters is such that exists a deterministic or stochastic bound on their backlog, i.e., the set of arrivals belongs to the stability region of the system; a mechanism that optimizes over the set of achievable rates and imposes proportional fairness among the D2D pairs performance. The works that are closer to ours are [34] and [33]. [34] computes the Pareto boundary of the stability region without considering user traffic

demands, but assuming that the transmission power of user devices can be tuned. Besides the fact that the analysis does not hold when the nodes are not in saturation, the paper applies to settings with few nodes. In our case, we assume that the power control mechanism acts on a different time scale than the scheduling of the resources. Nonetheless, our approach applies also to scenarios with large number of transmitters, and it remains accurate even when not all users are in saturation. The approach used by [33] uses a similar modeling of in-band underlay D2D communications via Coupled Processor System (CPS) modeling. Two major assumptions are used though. The scheduling policy taken into account is the Full Frequency Re-Use (FFR), i.e., D2D transmitters use the channel every time their transmission queue is not empty independently of the interference they cause. Furthermore, the buffers at the transmitters have finite, very limited capacity. As a result, the model studies a simplified version of the D2D system, where a monotonic CPS is derived from a Finite State Markov Chain characterization of the system. On the contrary, we generalize the study of D2D communications allowing buffers of large/infinite capacity and a complex, but much more realistic, scheduling policy. Such choice leads to a non-monotonic CPS characterization of the system, that could be analyzed, until now, just through worst case analysis.

Several pieces of work try to optimize the mechanism which assign resources to the different users when the system is in saturation. Those optimizations also hold when the aim is fairness. For example, [35] schedules resources in order to accomplish given Quality of Service (QoS) requirements, without taking into consideration the demands of the users while scheduling. Finally, [36] proposes a game theoretic approach for the scheduling of *each* single available resource (in time and frequency, i.e., the so called Resource Block (RB)). The parameters of the game are set in order for the transmitters to achieve a given fairness. Besides the complexity of the approach, that paper is designed to work only when all the devices have traffic to transmit.

2.1.2. Out-band D2D Performance Evaluation

Performance analysis of the out-band D2D communications had many more contributions in the state-of-the art. Mainly, this thesis differs from previous works because it presents an accurate approximation of achievable throughput, as well as a first approximation of airtime computation in D2D communications, in non-saturated scenarios. I.e., in scenarios where not all the users have always traffic to send. Furthermore, this thesis evaluates in details the complex interaction of relay out-band D2D communications in cellular networks.

Carrier Sense Multiple Access - Collision Avoidance (CSMA-CA) studies have traditionally focused on saturated traffic assumption [10]. Many of the available results for non-saturated conditions do not capture the effects of traffic dynamics on system performance. [37] for example, assumes that the probability of transmission and the probability of success of the stations in a non-saturated scenario is always the same during time, despite traffic dynamics do alter significantly such quantities. [38] expands the seminal work of Bianchi [10] to unsaturated conditions, including transmission errors and capture effects. Contrarily to what we propose in this thesis, all of these approaches are based on a Poissonian traffic assumption, and they depend heavily on the

assumption of traffic stationarity. [39] analyzes non saturation in heterogeneous traffic conditions, but its results still require a complete stochastic traffic characterization to be parametrized. This has been done despite of the fact that traffic in real networks is well far from being Poissonian (see [16] and related literature). In particular, traffic from live audio/video streaming exhibits a periodic behavior which substantially departs from the Poisson model, and which is characteristic of several known examples of instability [40]. This leaves open the issue of how to derive valid performance guarantees in out-band D2D communications in realistic settings, when little is known about traffic statistics. In the present work we propose a different approach, which assumes traffic to be constrained by *leaky bucket arrival curves* [14], which limit the maximum amount of bits which can arrive in a given time interval. As such a traffic characterization translates into quite loose assumptions on traffic statistics (mainly involving tail probabilities) and it applies to a large spectrum of practical settings.

2.1.3. CPS Performance Evaluation

Both in the case of in-band underlay D2D and in the case of out-band D2D, we based our analysis on the CPS model [12], which allows capturing the dependencies between user achievable rates due to sharing of the wireless transmission medium (mediated by the CSMA-CA mechanisms or the FlashLinQ scheduler) and traffic dynamics, which such coupling entails. CPS is a queueing system for which the service rate of each queue depends on the set of queues with ongoing job service [11–13]. CPS is a natural model for wireless communications since the interference experienced by users depends on the set of transmitting base stations. Chapter 4.2 and Chapter 5.2 show in details how the CPS model also fit for the particular cases of in-band underlay and out-band D2D we deal with in this thesis.

Due to its extensive possible applications, the CPS queueing system found interest in the research community. Nevertheless, the available results are very limited. [41] and [42] derive closed-form necessary and sufficient conditions for the stability of a CPS composed by just two queues with one class of traffic each, assuming Poisson arrivals and exponential service times. [11] derives a similar result, assuming heavy-tailed file size distribution. For a two-queue system with time-varying arrival rates, [43] develops a method for estimating the main statistics of the resulting two-dimensional queuing process.

For larger settings, some of the existing approaches are based on very conservative assumptions, which limit their scope and interest. For a specific inter-cell interference limited cellular networks scenario, [44] proposes an approximation method which assumes that the demands of all queues but one are smaller than the saturation rate. Results achieved by means of this approximation yield tighter accuracy when the system approaches saturation, but are loose in any other condition. For CPSs that are already known to be stable, [13] proposes a method for computing bounds on the moments of queue length, based on a semidefinite programming approach. Among the available results, those that apply to a generic number of queues rely also on the use of a large number of computationally heavy simulations for model parametrization, even under the Poisson

traffic assumption. For example, the analysis presented in [12], [45] provides a method to determine if a particular configuration of the CPS is stable, based on the steady state probabilities of underlying Markov chains. However, the derivation of the steady state probabilities requires a number of simulations which grows factorially with the size of the problem. In particular, if the system is composed by nodes, Markov chains must be analyzed to know if a given point belongs to the stability region. Furthermore, almost all existing approaches assume Poisson input traffic, leaving open the issue of how to derive performance bounds for CPS under different (and possibly more realistic) assumptions on input traffic. Notably, and most importantly, no existing results allow computing either soft (i.e., *stochastic*) or hard (i.e., *deterministic*) sufficient conditions for stability of the system, or soft/hard bounds on packet delay and backlog.

2.1.4. D2D-assisted Cellular Networks Performance Evaluation

The first part of this thesis concludes by deriving a novel comprehensive model which accounts for the intertwined nature of cellular and resource allocation when D2D relay is in place. In fact, while the bandwidth available at the cellular access affects the load of D2D over channels, it is also true that the performance of mobile terminals connected to Access Points (APs) and D2D transmissions are entangled and affect the quantity of traffic that can be offloaded. The framework introduced in this thesis allows for the computation of both achieved user-throughput and terminal consumption, even in non-saturated conditions. As of today, our contribution is the first presenting such features.

D2D-assisted cellular access networks have been proven to substantially improve throughput and fairness compared to standard cellular access networks and such a analytical tool may allow for the optimization of network operation. The works in [46–48] focus on D2D clustering in cellular networks. The authors of [46, 47] show via simulations that D2D clustering increases the network performance by up to in comparison to legacy cellular systems. [48] proposes an optimal resource utilization for multicast relaying with D2D clusters, provide a closed-form expression for the Probability Density Function (PDF) of the optimal number of relays in a cluster, and an intra-cluster retransmission scheme. [48] also show via numerical simulations that the proposed scheme achieves up to gain in resource utilization efficiency. The authors of [49] propose the so called Dataspotting approach that leverages clustering technique to increase the network capacity in dense networks. Dataspotting tries to avoid transmitting the same contents repeatedly to mobile users that are located in a small area. In Dataspotting, the operator maps the geographical availability of the contents in dense areas (i.e., data spots). When a user arrives to a data spot and requests for a content which is available in the data spot, the base station sends the information of the content owner to the user. Then the user initiates a WiFi communication with the content owner in order to retrieve the content. Their simulation results illustrate that there is a probability to find the users' desired content cached in the data spot.

Nevertheless, D2D-assisted cellular may cause an increase in power consumption [18, 19]. Existing network access techniques rarely account for energy efficiency and throughput at the

same time. They mostly aim to either improving battery lifetime of devices [50] or seek for throughput [51]. Moreover, energy efficient-aware approaches like [52] neglect the coupling between uplink and downlink resource allocation, or between cellular and utilization due to the operation of D2D relays. As of today there is no model or technique available to make informed access selection decisions based on analytical estimate of throughput and power consumption, and hence energy efficiency, in a mixed access network. Building on our model, we *formulate a new network-controlled access selection problem* that aims to maximize the energy efficiency achieved at terminal side. As a major point, in the proposed formulation, not only uplink and downlink are both simultaneously considered, but also they are jointly optimized in terms of cellular and resource allocation. Several techniques exists that deal with access selection and energy efficiency. [50] proposes to increase the battery lifetime of terminals by connecting them to the least power expensive technology among the ones available. No guarantees on the throughput achieved and no consideration on the different loads of the Points of Access (PoAs) is taken into account. Our algorithm, instead, dynamically reacts to the scenario of application, having the ability of detecting the load and the throughput achieved at each PoA.

Other proposals evaluate the most energy efficient access decision at the terminal, without using any central knowledge of the network situation. [53] assumes uplink and downlink demands are known at the terminal. The user then chooses the access technology that is able to serve its traffic with the lowest cost. However, in their scheme, access selection effects on neighboring terminals are not considered and terminals are not allowed to access the network if there is no PoA that can satisfy their demand. Furthermore, throughput and power consumption of users connected to the same PoA are tightly coupled. Our approach considers the impact of the terminal choice on all other terminals and allows PoAs to adapt and accommodate the requests of new terminals.

[52] accounts for D2D relay in downlink, though the energy efficiency of each terminal is considered as independent, and resource allocation as PoA-independent. Access selection is then solved as a knapsack problem. We overcome such limitations and propose a model that works at any system load and that accounts for uplink, downlink and for the coupling between resource allocation in and at the eNB, including in presence of D2D relay. [54], even if their work does not explicitly address an access selection mechanism, shows how, in an access network with in-band underlay D2D support, scheduling the least interfering users at once may reduce terminals power consumption, which is in line with our findings. However, we provide analytical reasons behind this phenomenon, which goes well beyond what hinted in [54].

2.1.5. Almost-Blank SubFrame (ABSF)

After having shown how to evaluate the performance of D2D communications both in-band or out-band, both as a direct connection among users or as a relay methodology, we show in the second part of this thesis how D2D communications can be managed together with other optimization techniques. The authors of [19] describe how D2D in unlicensed spectrum has a

higher implementation opportunity because it requires minor changes in the existing standards. In [55], it is shown that D2D can be deployed in Long Term Evolution (LTE) networks and groups can be formed with WiFi-Direct. The authors of [21] and [56] propose a practical opportunistic scheme and a protocol for D2D over LTE and WiFi networks. Following the conclusions of the above mentioned contributions, the second part of this thesis focuses on out-band D2D communications. Moreover, as in Chapter 8, D2D techniques are now considering not only WiFi-like options, but also novel and faster solutions, e.g., based on millimeter-wave communications [57].

In LTE the major source of interference is due to neighboring cell activity, a.k.a Inter Cell Interference (ICI). As we mentioned in Chapter 1, a common approach to mitigate ICI is to apply a 3GPP recently standardized mechanism, called ABSF [58]. ABSF coordinates different LTE evolved Node Bs (eNBs) by preventing their transmissions for a specified set of subframes (ABSF pattern), where only control signals are permitted. While D2D is attracting the attention of industrial players, ABSF has already become popular due to its trade-off between performance improvement and low implementation complexity, as widely shown by [59]. ABSF has been proposed for throttling macro base station transmissions in presence of micro and pico cells. However, much more interesting results have been shown when ABSF has been adopted for all kind of cells. Several works focus their research on how to properly design ABSF patterns to limit ICI and boost network performance, especially in high-density scenarios. The authors of [60] provide a clear overview about different ABSF proposals, classifying them as (i) semi-distributed, where a central entity coordinates scheduling resources, through ABSF patterns, while each base station is in charge of scheduling its users, and (ii) distributed, where each base station makes local decisions on its own ABSF patterns. As we show in this thesis, the first ones are the approaches which allow obtaining the best results. Only through the knowledge of the performance of all the cells in a given geographical area it is possible to choose the right portion of time that every eNB has to blank its transmissions. As semi-distributed ABSF mechanism, [61] presents BAse Station Inter-Cell Scheduling (BASICS). BASICS is an efficient algorithm that leverages ABSF to optimally increase the network throughput by serving best effort traffic and guaranteeing an acceptable level of fairness between users. Another semi-distributed ABSF mechanism is presented in [62]. In there, ABSF is tuned to guarantee inelastic traffic demand for delay-guaranteed networks. The authors of [62] tackle the ICIC problem by inspiring a heuristic solution which provides a near-optimal deterministic ABSF pattern to schedule all required traffic, when content distribution systems are involved. Another interesting solution, such as [63], deals with heterogeneous networks in which a macro base station coordinates the activity of small base stations to improve throughput performance when sharing a limited area. For a limited traffic distribution, the deterministic approach proposed in [20] aims to determine the best blanked subframe density according to a given traffic distribution. More advanced solutions focus on the pattern reuse which directly guides the ABSF activity pattern. In particular, [64] derives the best temporal pattern duration, given a set of chosen patterns to maximize the total user throughput. However, as proved in our work, while some scenarios may adversely impact on the system throughput, a

pure throughput maximization can lead to highly unfair throughputs. Finally, for what concerns the class of distributed ICIC mechanisms, [65] presents a lightweight fully distributed solution, where each base station makes its own scheduling decisions based on a game theoretic approach.

While both ICI reduction schemes and spectral efficiency enhancement solutions, as for example D2D, have shown promising outcomes, they have not been evaluated in a real-world scenario concurrently. The first work which evaluated, only via simulations, the impact of D2D (with WiFi Direct) and ABSF (with Base Stations Blanking (BSB)) was [66]. In [66], D2D relay speed is comparable to cellular speed, and the authors conclude that D2D and ABSF in combination can bring quite limited value added. However, very few analytical insights have been given before this manuscript on joint inter-cell interference coordination and intra-cellular traffic offloading.

Starting from the same assumptions, in Chapter 7 we present a novel mechanism, Two-level Opportunistic Spectrum Management for 5G Radio Access Networks (TOMRAN), which does not ignore the fact that WiFi can be a bottleneck for relaying cellular data. Although the achievable rate with WiFi is higher than with cellular technologies of the same generation, WiFi transmissions mainly suffer from poor coordination of transmitters, exhibiting high performance degradation. Our work differentiates from the related work because it takes into account the impact of different communication technologies on each other's performance, and dynamically adjusts offloaded traffic based on two factors: cluster size and set of achievable D2D rates, thus shedding light on how the performance is affected.

In Chapter 8, our work completely differentiates from the literature, since we are the first to analytically study and design a joint scheme to provide high spectral efficiency by leveraging cooperative D2D opportunistic communications using mm-waves connections, while at the same time adjusting user fairness by means of ABSF. The D2D physical layer is changed in order to remove the possibility that the intra-group resource sharing constitutes the bottleneck of the transmissions in the access network. Even though a compound analytic approach to D2D and ICIC is completely missing so far in the literature, many other solutions focus on the trade-off between throughput and fairness using different approaches. [67] and [68], for example, propose to apply user association and D2D multi-hop offloading, respectively, to achieve such a goal. Nevertheless, differently from our approach, many changes to the current cellular architecture should be done in order to implement such solutions.

As final note, recent studies on D2D communications show the feasibility of such schemes, including opportunistic scheduling requiring control decisions at the millisecond timescales [56].

2.1.6. Multi-Resource Sharing

The concept of multi-resource sharing, used in Chapter 9 has been widely studied over many years. Specifically, multi-resource sharing by flows with homogeneous requirements has been at the center of the efforts of the research community. Of particular interest for the present work is the observation made 30 years ago that network-wide max-min fairness is realized by implementing fair queuing in router queues and performing window-based flow control [69], [70]. This

claim was proved by [69] for a synchronous time-slotted network model with flow rates controlled by *hop-by-hop* windows [71]. However, there is no published proof, before this thesis, that max-min fair sharing occurs with *end-to-end* window control, as used in Transmission Control Protocol (TCP)/Internet Protocol (IP).

Our main objective here is to derive the equivalent result in the case of heterogeneous resource requirements, which also apply to the Centralized-Radio Access Network (C-RAN) access network described in Chapter 1. Explicitly, we show that imposing local fairness at each resource, i.e., assigning equal shares to each flow, coupled with end-to-end flow control results in a desirable generalization of max-min fairness called Bottleneck Max Fairness (BMF) [9]. For homogeneous requirements, the proof that backlogs eventually converge and rates stabilize at their max-min fair shares was the culmination of several years doctoral thesis work by Hahne [72]. [73] have since derived a somewhat simpler proof, thanks to their use of a fluid model, but this is still highly non-trivial and again confined to hop-by-hop window control. It is considerably harder to account for heterogeneous requirements because the water filling characterization of max-min fairness used in [71] and [73] does not generalize to BMF.

Many other multi-resource sharing technique have been presented. Ghodsi and co-authors introduced the problem of multi-resource sharing in compute clusters [74] and extended their analysis to networks [75]. They advocate so-called Dominant Resource Fairness (DRF). In networking applications, DRF requires schedulers at each resource to implement weighted max-min fairness with the same flow weight applied at each resource determined from the resource which requirement is maximal. This choice is motivated by a requirement that the allocation be *strategyproof*: flows should not be able to gain a greater bit rate by falsely stating their requirements. The plausibility of designing and implementing such a gaming strategy in a context of dynamic demand in a network setting is highly debatable, however. An alternative allocation that sacrifices strict strategyproofness in order to achieve a better efficiency–performance tradeoff, like BMF, may represent a better solution for nowadays networks. [9] showed that the BMF allocation always exists and has all the desirable sharing properties identified by Ghodsi [75] except strategyproofness. On the other hand, it has an alternative property called *single-bottleneck fairness*: if the network has a unique bottleneck, the allocation is such that each flow has an equal share of the bottleneck. That this property is not shared by DRF largely explains its inferior throughput performance. We argue that this difference of performance outweighed the lack of strategyproofness, especially in an environment where cheating is technically impossible (like wireless) or practically impossible (like C-RANs).

In the wired Internet, bandwidth sharing is realized by means of congestion control protocols like TCP that react to drop signals received from First-In First-Out (FIFO) buffers. Sharing is generally fair enough if users implement the same protocol [76] though it has often been noted that fair queuing implemented in router queues would provide more robust control, e.g., [77], [78]. For a wireless link, where requirements are highly variable, it is generally considered preferable to aim for equal resource shares (as in BMF), rather than equal bit rates. This is broadly what the

proportional fair scheduler achieves [79], as implemented in 3G and 4G cellular networks. That the b scheduler tends to realize max-min fair *bit rates* was recognized as a performance anomaly [80]. This anomaly is mitigated in more recent WiFi versions, where packet aggregation indeed tends to equalize resource shares. In C-RAN, the dynamic provision of compute capacity implies CPU may become a temporary bottleneck and the way it is shared is therefore an issue. Simple FIFO queuing coupled with end-to-end congestion control would lead to approximately equal flow bit rates, as in a wired Internet. However, if requirements differ significantly, max-min fair rates would produce the same "performance anomaly" as in b. In an early paper considering dual (CPU and bandwidth) resource sharing, [81] proposed an Explicit Congestion Notification (ECN) marking scheme intended to realize proportional fairness through TCP congestion control. We would argue that the use of a scheduler to equalize CPU usage among flows constitutes a more satisfactory resource sharing solution.

[82] shows that C-RANs are able to compensate for unbalanced computational loads at the base stations. Each base station, indeed, would have to code and decode the traffic of the users attached. The amount of computational resources needed would be varying depending on the movement and traffic patterns and centralizing the computational efforts in a datacenter would prevent from over provisioning. Nevertheless, [82] overlook the problem that a centralized datacenter may create a bottleneck to the end-to-end communications, since the results do not characterize, if not numerically, the fluctuation of the computational resources requirement due to channel parameters and user behavior. Finally, [8] proposes a scheduling methodology for C-RAN, which simultaneously deals with computational and wireless resources. The approach assigns the computational resources of the datacenter ordering the different transmissions by spectral efficiency up to saturation. Such a decision may create severe fairness performance issues in case of heterogeneous scenarios, where users experience very different channel performance.

2.2. Network Calculus Basic Concepts and Definitions

In this section, we introduce some definitions and basic results used in our analysis (especially in Chapter 3). NC [14] is a min-plus system theory for deterministic performance analysis of a queuing system. It provides tools for the derivation of bounds to backlog and packet delay in a network. Traffic in NC is typically characterized by means of *arrival curves* [14].

Definition 1 (Arrival Curve). *Let be the set of nonnegative wide-sense increasing functions. , let be the cumulative traffic arrival function for the time interval . Then is an arrival curve for the considered flow if for any , .*

As such a traffic characterization translates into quite loose assumptions on traffic statistics (mainly involving tail probabilities), it applies to a large spectrum of practical settings. For instance, any traffic source on the Internet is constrained by some form of arrival curve (e.g. due to limitations of the application generating the traffic, and/or to the bit rate of the connection), possibly by means of some conservative assumptions on the statistics of the generated traffic.

One of the most common classes of arrival curves is the *leaky bucket* arrival curve, where . The non-negative parameters and are, respectively, the *leaky bucket rate* and the *burstiness* of the traffic flow. Leaky bucket arrival curves are well suited to capture the bursty nature of traffic flows in data networks [16, 83] and they are part of the DiffServ policing mechanism [84].

Let indicate the backlog of queue at time . In what follows we will adopt the following two notions of stability, according to whether traffic is characterized stochastically or through arrival curves.

Definition 2 (Deterministic stability [14]). *A system of queues is deterministically stable if there exists a such that*

(2.1)

Definition 3 (Stochastic stability [85]). *A system of queues is stochastically stable if there exists a such that*

(2.2)

Note that deterministic stability implies stochastic stability, but the converse is not true. Practical sufficient conditions for stability typically imply some constraints on source traffic (e.g., on arrival statistics, or on their arrival curve parameters) and/or on the network (i.e., on some form of service guarantees at queues) [86]. For instance, for queues in isolation, a sufficient condition for stochastic (respectively deterministic) stability is that at each queue the average arrival rate is less than the average service rate (respectively, that the leaky bucket rate for the arrivals is less that the guaranteed service rate) [85]. Other network calculus concepts we will use in our analysis are the *continuous data scaling block*, the *policer* and the *multiplexing block*.

Definition 4 (Continuous Data Scaling block). *Let denote the cumulative arrivals to a node up to time . The node is a continuous data scaling block, with scaling value $\mathbb{R}$, if for any time interval , the amount of served traffic is .*

Scaling blocks have been first introduced to model transformation processes which alter the total amount of traffic (lossy channels, data processing, encoding/decoding, discard of non-conformant traffic [87]).

Definition 5 (Policer). *A policer with rate is a processing node that, for any arbitrary input traffic, forces as the maximum departure rate at its output.*

This definition implies that the policer forces as an arrival curve for the output traffic. In what follows we consider *unbuffered* policers, which discard non-conformant traffic. If denotes the cumulative arrivals at the input to the policer up to time , this implies that for any time interval the output of the policer is .

Definition 6 (Multiplexing block). *Let denote the cumulative arrivals to a node up to time . A multiplexing block, with multiplexing factor , is a node with output flows, such that the cumulative arrivals at the -th output are the same as at the input, i.e.*

.

Finally, we recall the concept of Generalized Processor Sharing (GPS) node [88]. In general, a GPS node is composed by a server having a fixed service rate , and different queues. Each queue is characterized by a weight , with . At any given time, is split among the non-empty queues proportionally to their weights. At the fluid limit, if queue is active at time , the service rate reserved to queue is:

(2.3)

where is the set of active queues at time .

Part I : An Analytical Framework for D2D Performance Evaluation

Chapter 3

A CPS Model for D2D Communications

In this chapter we consider a system composed by a network of queues. The service rate of each queue, at any given time, depends on the set of queues with ongoing job service. This type of model is commonly referred to as Coupled Processor System (CPS) [11–13], and it arises in several contexts in which coupling derives from resource sharing.

The CPS has been also proposed to model the complex interdependence of wireless transmissions' performance [89]. In wireless networks, the interdependence of transmissions' performance is typically due to interference, since devices share the same medium. In order to understand why the CPS is a good model for wireless networks, it is sufficient to think of a scenario where several transmitters may occupy simultaneously the same spectrum. In this case, the interference sensed at destination depends on the particular set of simultaneously active transmitters. Since transmitters use a particular Modulation and Coding Scheme (MCS) depending on the interference sensed at destination, the set of simultaneously active transmitters dictates somehow the "service rate" at which each transmitter serves its own traffic. This aspect is perfectly captured by the CPS modeling, since CPS's queue service rate depends on the set of simultaneously active queues. CPSs are therefore an essential tool in the evaluation of wireless systems performance, especially when Device-to-Device (D2D) communications are in place. As said in Chapter 1, indeed, if D2D communications are in place, having more than a transmission happening on the same spectrum is far from being rare.

Unfortunately, CPS performance analysis is as complex as wireless performance analysis. Up to now, the characterization of the delay experienced by the CPS's arrivals, as well as conditions on traffic statistics which are able to ensure queue length stability, are known only for a 2-queue CPS. Therefore, the use of the CPS modeling has been limited to toy examples. In the following, we propose a fully analytical approach to the performance analysis of a generic -queue CPS, which does not require simulations to be parametrized. The proposed method is independent of the nature and shape of coupling relations. Furthermore, our method can be applied to systems in

which the traffic is described either stochastically or by means of deterministic arrival curves [14].

The main results presented in the following chapter are:

- We present a *generalized* method for the derivation of performance bounds in a CPS with an arbitrary amount of queues, based on a conservative characterization of the service behavior of a CPS.

- By applying our method to CPSs with independent and stationary arrivals, as well as to input traffic constrained by arrival curves, we derive analytical expressions of sufficient conditions for stability of the system, and for bounds on delay and backlog. For arrivals described stochastically, state-of-art queueing theory results are exploited. If instead arrivals are described by arrival curves, our analysis exploits Network Calculus (NC).

The rest of the chapter is organized as follows. Section 3.1 and Section 3.2 present, respectively, the system model and the idea behind the method we use to assess the performance of a CPS. Section 3.3 describes in details our method for the derivation of performance bounds in a CPS in which service rate at each queue decreases with the size of the active queues. Section 3.4 illustrates the derivation of analytical expressions for the main performance bounds. Section 3.5 extends the proposed approach to queue systems with generic, not necessarily monotonic, coupling.

Notice that we show how to apply the model to D2D communications in Chapter 4. In Chapter 4 we also assess numerically the tightness of our bounds to the real performance of a CPS, when the CPS is used to model an in-band D2D scenario.

3.1. Model and Assumptions

3.1.1. System Model

We consider a system of queues, where each queue receives traffic from one or more fresh sources (i.e., sources residing out of the system). We assume such queues are served by work conserving schedulers, and that the served traffic leaves the system. The system *state* at time , denoted as , is an array of binary variables, such that for each , the -th variable is if the th queue have traffic to serve, and otherwise. We assume that at any time , the service rate of queue is determined only by the system state, i.e., . Therefore, each traffic flow is served at the fluid limit, experiencing different service rates depending on the system state. Nonetheless, the service rate of a particular queue can assume only a fixed number of different values, one for each of the possible system states. This represents the formal definition of CPS.

Initially we tackle the performance analysis of *monotonic decreasing* CPSs. That is, if

and are two different states such that [1], then . This class of CPSs includes many problems of practical interest, e.g., in the context of wireless networks and of bandwidth sharing in packet networks [12]. Indeed, resource sharing generally induces coupling which satisfies the monotonic decreasing property, as usually the more contend for the same resource (e.g., bandwidth in a wireless network), the less is the share that every contender gets. However, in some relevant practical settings such coupling might significantly deviate from monotonicity. For such cases, which have been little studied so far, Section 3.5 will extend the approach to a generic CPS.

Finally, we consider that no losses occur in the system (i.e., queues have infinite capacity).

3.2. The "Upper Bounding" Approach

In this section we outline a new approach to the analysis of CPSs. It is based on the concept of *upper bounding* network.

Definition 7 (Upper bounding property). *Given two networks of queues, we say that one network upper bounds another when the stochastic (risp. deterministic) stability of the former implies the stochastic (risp. deterministic) stability of the latter.*

Please note that the definitions of deterministic and stochastic stability have been presented in Section 2.2. When a network upper bounds another, then sufficient stability conditions for the upper bounding network hold also for the upper bounded network.

Our approach consists in individuating at least a network which upper bounds the CPS we want to study, and for which we are able to compute sufficient conditions for stability. In order to overcome the difficulties of existing approaches to CPS analysis, we will focus on upper bounding networks which allow applying methods of analysis that are not directly applicable to the original CPS, such as standard NC or simple Queuing Theory principles. We provide a method to construct a class of upper bounding networks satisfying these properties, and we derive stochastic and deterministic bounds holding for the upper bounding networks. By exploiting the upper bounding property, from stability conditions and bounds valid for the given class of upper bounding network we derive performance bounds for the upper bounded CPS. The bounds obtainable with this approach of course depend on the specific features of the upper bounding network. Therefore, for a CPS, our method yields as many bounds as upper bounding networks. In what follows we describe a first application of our approach. In particular, for a given CPS, we illustrate a class of upper bounding networks that allows deriving new CPSs performance bounds by using either NC or queueing theory.

[1] Note that the *component-wise* inequality implies that the set of queues active in state is a subset of the set of queues active in state .

3.3. Monotonic Decreasing CPSs

In this section, we apply our approach to *monotonic decreasing* CPSs. Given an -queue CPS, we start by illustrating a fundamental property for a network to upper bound the CPS. Then, we present a class of feed-forward networks, that we denote as *ancillary networks*, which respects the above mentioned property. Finally, from stability conditions and performance bounds holding on ancillary networks, we show how to derive stability conditions and bounds for their associated CPS.

3.3.1. A Fundamental Property for Upper Bounding Networks

We start by presenting a result defining a sufficient condition for a class of networks to upper bound a CPS. The structure of the ancillary networks we present in Section 3.3.2 is based on the following result.

Lemma 1 (Upper Bounding Network). *Consider a network with queues, such that there is a one-to-one mapping between the queues of the CPS and a subset of queues of the network. If the queues of the subset are labeled with and the queues of the CPS are labeled by , the mapping is such that each queue of the subset has the same arrivals at any time as its corresponding queue in the CPS. Let and be the service rates at time , respectively, at queue and queue . If, at any time , it holds that for each queue of the CPS, then the network upper bounds the CPS.*

Proof: We prove that at any time , the length of each queue of the CPS is always smaller than the queue length of the corresponding queue in the network. If this is the case, a bound on the backlog of also holds for the corresponding queue of the CPS, so that the partial stability of the network implies the stability of the CPS. We prove that by contradiction. Assume that is the first time for which holds. As arrivals are the same in both queues, this implies that at time a bit of traffic has left queue in the network, while is still being served at the corresponding queue of the CPS. Assume the last bit served before at queue has been served at in the network, and the last bit has been served at queue at time . From the definition of , . This means it exists an instant , at which , contradicting the assumptions.

Basically, Lemma 1 says that a network upper bounds an associated CPS if each queue of the network serves its traffic at a rate slower than the one used by the mapped CPS queue.

3.3.2. Structure of an Ancillary Network

We now present a class of networks, the ancillary networks, which can be used to upper bound a monotonic CPS. Notice that the presented structure has been thought with the aim of respecting Lemma 1.

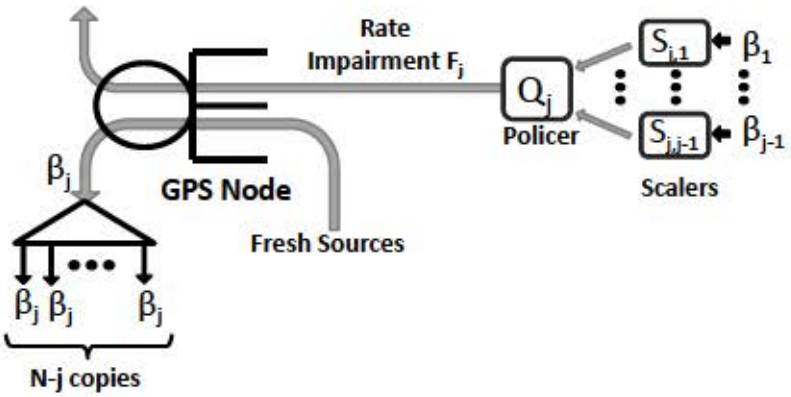


Figure 3.1: The -th stage of a -queues ancillary network.

The topology of an ancillary network is feed-forward, with stages and a two-queue work-conserving Generalized Processor Sharing (GPS) node at each stage. GPS nodes work at the fluid limit and are in a one-to-one mapping relation with the queues of the CPS. Let indicate one of the possible permutations of the labels of the queues of the CPS. Each permutation individuates a specific mapping, which associates the -th queue of the CPS to the GPS node at the -th stage of the ancillary network, . At any time , we assume arrivals from fresh sources at stage are the same as at the corresponding queue at the CPS.

Let us consider the structure of the -th stage (Fig. 3.1) of an ancillary network with mapping . By we denote the service rate of queue of the CPS when the active queues at the CPS are . is instead the *saturation rate* of , i.e., the service rate of queue of the CPS when all queues are active. Each stage receives traffic from the previous stages. We denote as the **arrival rate** from stage to at time . Each of these flows feeds a scaling block, with scaling coefficient . The aggregate output of all the scaling blocks at stage then feeds to a policer with rate . The output of the policer feeds a dedicated queue of the GPS. The other queue of the GPS, instead, serves traffic from fresh sources. The total capacity of the GPS node is , and the GPS weights are for fresh traffic, and for traffic from the policer. This means that the guaranteed service rate for fresh traffic is , while the guaranteed service rate for the traffic coming from the policer is . At the output of the GPS node, the traffic from the policer exits the network. The traffic from fresh sources instead feeds a multiplexing block with factor . Each of the output flows then feeds one of the following stages Fig. 3.2 illustrates the structure of an ancillary network for .

It can be easily seen that this choice of GPS weights ensures that the minimum service rate for fresh traffic is equal to the saturation rate of . That is, each GPS node serves the fresh traffic at least with the minimum rate guaranteed by the CPS. This particular choice of GPS weights

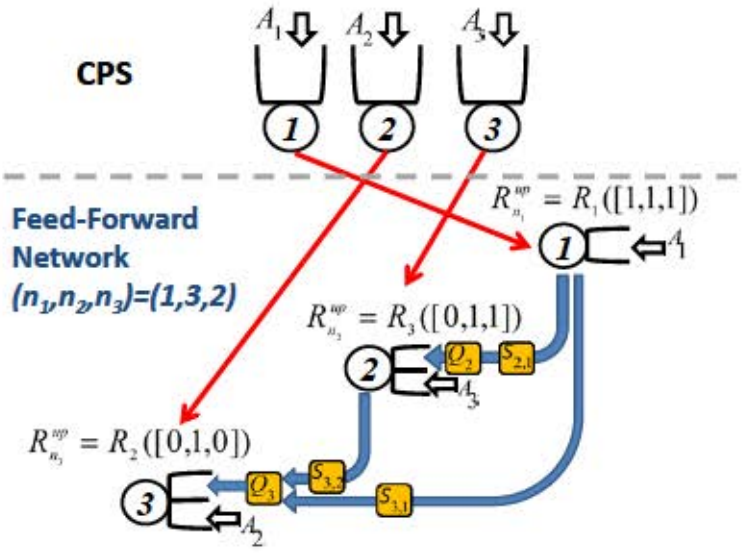


Figure 3.2: An example of a three-queue CPS, and of an ancillary network associated to it, corresponding to the mapping 1 , 2 , 3

also ensures that the guaranteed service rate of the traffic coming from the policer is equal to its maximum arrival rate. As a result, the traffic coming from stages does not buffer at the GPS node and its presence at the dedicated queue is only dictated by the presence of traffic to serve at queues for fresh traffic at stages .

When present in the dedicated queue, the traffic coming from the policer absorbs part of the capacity of the GPS node. Therefore, we call the flow coming from the policer *impairment flow*. The service rate of the impairment flow, which we denote as , is given by

$$\left(\sum_{=1}^{1} \quad , \quad , \right) \tag{3.1}$$

The in the computation of is due to the presence of the policer that precedes the dedicated queue of the GPS node. Indeed, the second argument of the minimum represents the maximum output rate allowed by the stage policer, while the first argument is the incoming rate at the policer, i.e., the summation of the rescaled flows coming from the preceding stages.

The role of the scalers is to allow tuning the incoming traffic at each stage. As we show in the following section, there always exist at least a set of values of the scaling coefficients such that the service rate for fresh traffic at the -th GPS node, i.e., , is never larger than the service rate at the corresponding CPS queue . This observation is the key idea behind the structure of the proposed ancillary networks, and it allows to satisfy Lemma 1.

Let be the matrix of the scaling coefficients of an ancillary network. Hence, given a CPS, an ancillary network associated to it is univocally identified by the pair .

3.3.3. Sufficient Conditions for the Upper Bounding Property

We start presenting an important property of the ancillary networks. As we mentioned earlier, at each stage , the service rate offered to fresh traffic at the GPS node is . Hence, such service rate depends on the activity of nodes in stages . Specifically, in the following Lemma, we show how univocally depends on the state of the queues of the GPS nodes dedicated to fresh arrivals.

Lemma 2. *Let be an array of N binary variables, such that the -th variable is if the queue for fresh arrivals at the GPS node of the -th stage of the ancillary network is busy at time , and otherwise. Then for each , **is univocally determined by** , i.e.,*

.

Proof: We prove the lemma by induction. We start from the first stage of the ancillary network. in case there is traffic from fresh sources in the dedicated queue, 0 otherwise. As a result, , that is, if , 0 otherwise. univocally depends on the state of . Since there is no dependence of on the state of the remaining following stages, we can say that depends, more generically, on . Let us now assume that for the first stages, the lemma holds. I.e, . Therefore, , , is univocally determined by , i.e, . From Eq. (3.1), is then obvious that can assume only a finite number of values, one for each possible value of .

In what follows we exploit Lemma 2 to derive a set of sufficient conditions on the instantaneous rate of the impairment flow (and therefore on scaling coefficients) that an ancillary network must satisfy in order to upper bound the CPS.

Theorem 1. *It is given a monotonic decreasing CPS and an ( ) ancillary network associated to it. The ancillary network upper bounds the CPS if ,*

(3.2)

where represents after applying the permutation . In practice, is the service rate of queue of the CPS when the queues active at the CPS are exactly the ones mapped onto active queues of the ancillary network.

Proof: Let us indicate with the service rate for fresh traffic at the -th GPS node. Then Eq. (3.2) can be written as

(3.3)

Eq. (3.3) ensures that the service rate of each queue for fresh traffic of the network is always inferior to the one of the correspondent queues in the CPS, if all active queues for fresh arrivals in the ancillary network are the only active queues in the mapped CPS. Nevertheless, the set of

active queues of the CPS is , not . The condition in Eq. (3.3) would be sufficient if the set of active queues in the CPS was the same of the ancillary networks at any time . Given the fact that the CPS is monotonic, the only possible way for fresh arrivals at the network to be served at a faster rate than at the CPS is that at least one of the queues in is empty in the corresponding queue of the network at time , i.e., does not belong to the active queues in .

We now prove that Eq. (3.3) is sufficient for the ancillary network to respect Lemma 1, even taking into account that . That is, we prove that if Eq. (3.3) holds at any node and any time , then for any node and any time . If so, from Eq. (3.3) and thanks to the monotonicity of the CPS, it follows that

hence proving the theorem.

In the following, we prove that by contradiction. Considering that the arrivals are the same at the nodes of the network and at the corresponding queues of the CPS, we can assume that at , . Then, we assume by hypothesis that is the first moment at which a queue that is not empty at the CPS turns empty in the corresponding stage of the ancillary network. In other words, is the first moment at which does not hold. In order to satisfy this hypothesis, it exists at least a at which . For what previously said, this is possible just if at one of the queues for fresh traffic is empty at the network, while the corresponding queue at the CPS is not. The existence of contradicts the definition of , proving the theorem.

The proof shows how, as a consequence of Eq. (3.2), at any moment and at any node of the network the service rate for fresh traffic is never larger than the rate at the corresponding queue of the CPS. From the structure of the ancillary network, it can be easily verified that for every mapping there are at least two choices of scaling coefficients which always satisfy Theorem 1.

Theorem 2. *Given an -nodes CPS, for every mapping there exist scaling matrices , , with*

- , ;
- ———, ,

such that the ancillary networks ( ) and ( ) satisfy Theorem 1, i.e., they upper bound the given CPS. is the service rate at the CPS queue when the CPS queues are active.

Proof: In order to satisfy the condition in Eq. (3.2), we should prove Theorem 2 for each . Nevertheless, by its definition, , which both sides of Eq. (3.2) depends on, takes only a finite number of values. Specifically, can assume values, one for each of the possible combinations of the binary states of the queues for fresh arrivals at the ancillary network.

For the first choice of the scaling values, substituting in Eq. (3.2), for any value of . Since in a monotonic CPS, Eq. (3.2) is always respected.

For the second choice of the scaling values, we assume that is the first element in equal to . In this case:

(3.4)

is indeed equal to , since is the first stage with traffic to serve at the fresh arrivals queue and no traffic is entering from previous stages at stage . Depending on , the right hand-side of Eq. (3.2) will assume some fixed value. In particular, it will be equal to . Since considers all binary variables equal to 1 by its definition, and thanks to the monotonicity of the CPS, . Therefore:

and Eq. (3.2) is respected. Since this reasoning applies to any (and therefore to any time , which makes vary), this proves the theorem also for the second set of presented scaling values.

3.4. Sufficient Conditions for Stability and Bounds on Backlog and Delay

In the previous section, we have seen how to derive a set of upper bounding networks for a given monotonic decreasing CPS, i.e., the ancillary networks. In what follows we present some sufficient conditions for stability of a CPS, as well as bounds on packet delay and backlog, for two large classes of traffic characterizations. Results are obtained by exploiting the feed-forward property of the ancillary networks.

3.4.1. Independent and Stationary Sources

Here we consider the case of a monotonic decreasing CPS with stationary and independent arrival processes, with average arrival rates (in bits/s), . For such arrival processes, the following theorem yields sufficient conditions on average arrival rates for the CPS to be stable.

Theorem 3. *It is given an -queue monotonic decreasing CPS with set of average arrival rates for its inputs equal to . Sufficient conditions for the CPS to be stochastically stable are that it exists at least an ancillary network satisfying Theorem 1 and such that*

the arrival rates satisfy:

(3.5)

where is the mean rate impairment at stage . is:

(3.6)

is the sub array of and depicts the state of the stages preceding .
is the probability for the first stages of the ancillary network to be in state and
represents the average over time. is given by:

(3.7)

Proof: By applying Bayes' result, and exploiting the independence between arrival processes, we have that ,

(3.8)

where is the probability that the queue at the -th stage is in state . We now prove Theorem 3 by induction.

First step: At the first stage, fresh traffic is served at a constant rate equal to , as there are no external arrivals from previous stages, i.e, endogenous arrivals. Hence the sufficient condition for stability of that queue is . At the second stage of network , when the queue for fresh traffic at the first stage is active, the arrival rate of traffic from the first stage is . Hence, from the structure of stage 2, the traffic in input to the policer is given by . As the policer at stage 2 has policing rate , the mean impairment rate at stage 2 is given by

(3.9)

is equal to the mean utilization at node 1, which is given by ——. Finally, by the node stability condition at the queue for fresh traffic at node 2, the mean arrival rate must be not larger than the mean service rate for fresh traffic at that node. Given that the GPS node works at the fluid limit, the service rate for fresh traffic is equal to the minimum service rate dedicated to fresh arrivals, , plus the part of service which is not consumed by the endogenous traffic. I.e., . We have therefore that .

Induction step. We first note that for a given , the arrival rate at the policer of stage is . Hence, the rate impairment at stage , for a given , is the min between the arrival rate at the policer and the policing rate. The average rate impairment at is then computed by Eq. (3.6). In order to compute the probability that the first queues are in state , we can compute as ———, i.e., as the average utilization of the queue for fresh arrivals at stage . , instead, is simply . Recalling Eq. (3.8), this result confirms Eq. (3.7) for stage . Since the GPS works at fluid limit, the average service rate reserved to fresh arrival traffic is plus the fraction of capacity reserved for the endogenous traffic that remains unused, i.e., . Please note that is always less than due to the policer at stage . By imposing node stability at the fresh traffic queue we get the expression of Eq. (3.5).

These sufficient conditions are derived by analyzing the upper bounding network by stages, and imposing node stability at each GPS.

It is generally hard to compare our sufficient conditions to existing results, as (except for those assuming saturation in the system) they rely on simulations. For simple 2-queue CPSs, nonetheless, analytic expressions are available [12]. Our conditions for are:

—

—

Such expressions, for the choice of scaling values described in the previous section (i.e., ——— where is in the first case and in the second) are identical to those derived in [12]. Hence for simple cases, our sufficient conditions comply with the conditions of other approaches, and in particular with [12], whose conditions (valid only for Poisson arrivals) are also necessary. Please note that , since at the first stage the other queue mapped onto the following stage is considered as always active.

When the sufficient conditions in Theorem 3 hold, it is possible to derive mean backlog and delay by exploiting standard queuing theory results. For instance, when arrivals are Poisson and packet lengths are exponentially distributed, the fresh arrivals queue at any stage of the upper bounding network can be approximated as an queue. From the Pollaczek-Khinchine formula, the total traffic in the queue at stage , i.e., , is given by:

—

——————— (3.10)

where is the mean queue utilization, given by ————, is the packet processing time for fresh arrivals at the -th stage and is the mean packet size.

can be computed as . Exploiting the law of total expectation:

is derived assuming the node has exponential service times, with mean ————.

is the mean rate impairment at stage when the state of queues to is given by , and it can be derived from Eq. (3.1). The derivation of follows the same steps. Furthermore, mean packet delay is derived from by using Little's law.

3.4.2. Traffic Constrained by Arrival Curves

In what follows, we assume fresh traffic is constrained by arrival curves. More specifically, we consider leaky bucket arrival curves, despite similar results can be derived through our method for other types of arrival curves. The following theorem defines a set of sufficient conditions on leaky bucket rates for the deterministic stability of the CPS.

Theorem 4. *Given an -queue monotonic decreasing CPS, where at each node fresh arrivals are constrained by leaky bucket arrival curves, with parameters , the CPS is deterministically stable if there exists at least one associated network satisfying Theorem 1, and such that at each stage , satisfies*

(3.11)

Proof: Theorem 4 can be proved by induction on the index of the stages of the feed-forward network. Let us consider a network ( ) associated to the given CPS, and satisfying Theorem 1. Let us also consider stage . As stages to are stable, , is constrained by a leaky bucket arrival curve with rate . Hence, from Eq. (3.1), the rate impairment is constrained by a leaky bucket arrival curve, with rate equal to . Eq. (3.11) derives from imposing that the sum of the leaky bucket rates of all arrivals at the GPS node should be less than its total service rate .

From Eq. (3.11) we see that a trivial sufficient condition for stability is , , which corresponds to assuming the whole system is in saturation (all queues are always active). The additional terms in Eq. (3.11) take into account the fact that queues are not always active. The additional terms are a function of the bounds to fresh traffic at those nodes which affect the service rate of the considered stage.

When a CPS is stable according to Theorem 4, the set of upper bounding networks can be also used to compute hard bounds to delay and backlog for the CPS.

Theorem 5. *Let us consider a monotonic decreasing CPS and an upper bounding network which verifies the hypothesis of Theorem 4. Then a bound for backlog at each node of the CPS, is given by:*

(3.12)

Moreover, if the nodes of the CPS are FIFO, a bound to packet delay at the same node is:

(3.13)

with:

(from Eq. (3.11))

Please note that is infinite just if . Therefore, there is no configuration for the arrivals that is deterministically stable leading to unbounded delays. The derivation of the bounds in Theorem 5 is a straightforward application of elementary NC results, and is therefore omitted.

3.4.3. Bounds Across Multiple Upper Bounding Networks

All the sufficient conditions for stability as well as all the bounds presented so far are based on the individuation of a specific upper bounding network, i.e., of a given ordering of queues and of a given choice of scaling factors which satisfy Theorem 1. Each upper bounding network leads to a set of bounds generally different than those obtainable with other upper bounding networks (i.e., associated with a different pair).

Hence the ultimate set of arrival rates which leads to stable backlogs at the original CPS, and that can be obtained with our approach, is given by the union of the set of "stable" arrival rates obtained through all the upper bounding networks satisfying the hypothesis of Theorem 3 or Theorem 4 (depending on the type of arrivals). Similarly, for a specific set of inputs, the ultimate bounds to backlog and delay can be derived as the across all bounds derived from upper bounding networks which, for those inputs, satisfy Theorem 3 or Theorem 4.

In practice, as there are potentially infinitely many of such networks, strategies are required

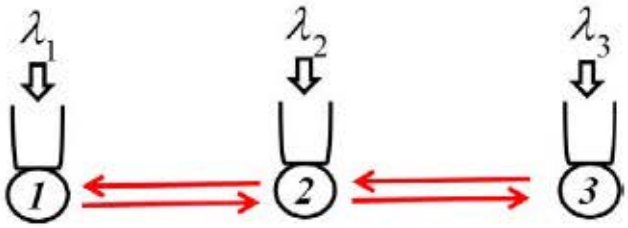


Figure 3.3: Example of a CPS exhibiting non-monotonic behavior.

to efficiently restrict the search to a small subset of networks. Such strategies are often based on characteristics of the specific system under study. In this thesis, we apply the CPS modeling to out-band and in-band D2D communications. Chapter 4 and Chapter 5 will therefore show how to optimize the search over a small number of networks, given the particular case in hand.

3.5. Analysis of a Generic CPS

So far we have assumed all CPSs to be monotonic decreasing. Such assumption applies to a large set of practical problems modeled as queuing systems with coupled processors. However, in several cases, some of the systems to which the CPS model has been applied in the literature exhibit a non-monotonic behavior. For instance, in Chapter 4 we model in-band D2D communications only by means of non-monotonic CPSs. A generalization of the results presented in Section 3.4 is therefore needed to account for any generic CPS.

In real systems, non-monotonic behaviors are typically found in CPSs which are generally monotonic decreasing, but in which, for some pairs of processors, the interdependency is not direct but mediated by the activity of a third queue.

For example, let us consider three coupled queues such as the one in Fig. 3.3, in which the arrows indicate the coupling relationship. In such example, there is no direct coupling between queues and . Recalling that the state of such a CPS is a binary triplet, where the -th position is 0 or 1 according to whether the -th queue is busy or idle, let us assume the following relationships between the service rates of the three queues:

In such relationships between service rates, the non-monotonicity is in the relationship and . Such a CPS might model, for instance, a set of wireless transmitters in which the activity of transmitter (resp.) prevents transmitter from interfering at queue (resp.).

In this example, let us consider the case of stationary and independent arrival processes, as

considered in Section 3.4.1 for monotonic decreasing CPSs. Note that the example can be easily extended to leaky bucket arrivals. Let us consider the maximum arrival rates for which the system is stable in the following two configurations. In the first one, . Then, queues and constitute a monotonic decreasing CPS. As we have seen in the two-queue example in Section 3.4.1, a stable allocation of arrival rates is , and . Such allocation lies on the border of the stability region of the system, i.e., we cannot increase the arrival rate at one queue without decreasing the arrival rate at the other, if the system has to be kept stable. In the second configuration, let us assume the same arrival rates at queue and , and let us consider the case in which queue is always busy. Again we can derive the stability region of such CPS by reduction to a two-queue CPS. In this way we see that is still sustainable by queue , but now queue could be stable with a larger arrival rate than the one of the previous configuration ( ), hence with a larger arrival rate than when queue was always idle.

In what follows, we discuss how to extend our approach to the case in which the CPS is not necessarily monotonic decreasing. Such extension is based on the fact that in a generic CPS at each queue the smallest service rate does not necessarily correspond to the case in which all queues are busy, nor does the largest service rate correspond to the case in which all queues are idle.

3.5.1. Feed-forward Upper Bounding Networks

In a generic CPS, the general structure of the feed-forward network that we propose in order to derive performance bounds is the same as described in Section 3.3.2. However, the values taken at stage by the total service rate of the j-th GPS node have to be adjusted to the non-monotonic nature of the coupling. This translates into a generic procedure to find GPS weights and policy parameters. We recall that the upper bounding feed-forward network is derived from the original CPS by breaking some of the circular dependencies between queues in a way that results only in feed-forward dependency. Any node at a stage of the feed-forward network is influenced only by GPS nodes from earlier stages, and influences only nodes at later stages. For the resulting network to be upper bounding the original CPS, the effects of coupling of those stages whose influence is not modeled explicitly by exchange of traffic must be factored into the service rates of all GPS nodes in a conservative way. In a monotonic CPS, this conservative approach considers all the following stages (whose coupling relationship is not modeled via exchange of traffic) as always active, i.e., it assumes that the coupling effects are always present, independently from traffic. In a non-monotonic CPS instead considering all the following stages as always active may not represent a conservative approach. In order to cope with this difference, when it comes to non-monotonic CPSs, the main modifications to the ancillary networks presented in Section 3.3.2 are to the GPS node capacity and to the policing rates.

When we set the capacity of the GPS our goal is to consider all the following stages as always active. Let be the state vector for the queues to of the CPS, and let be a vector of

all zeros of length representing the state vector for the queues to of the CPS. If we reorder the labels of the CPS queues following the particular sorting , the vector state of the CPS is , where is the binary state of the CPS queue . We can extend the definition of as follows:

(3.14)

Given the new definition valid for non-monotonic cases, models in a conservative way the interference at stage of all the stages which follow . is equal to the "worst case" service rate at the CPS queue (computed over all the possible states of queues to) when queues to are idle. In the monotonic decreasing case this definition coincides with the one in Section 3.3.2, since in that case the minimum service rate is achieved when queues to are all active.

In order to adapt the policing rate and the GPS weigths to the non-monotonic case, we change similarly . In the case of generic CPS, is the worst case service rate of queue in the CPS over all -dimensional system states , i.e.:

(3.15)

The GPS weights for the fresh traffic at stage is for fresh traffic, while it is for traffic arriving from the policer. The policing rate is equal to . and are as in Eq. (3.15) and in Eq. (3.14), respectively.

We recall that represents the binary vector state at time of queues for fresh arrivals at the ancillary network obtained applying the permutation to the CPS queues. In order to generalize Theorem 1, we modify in Eq. (3.2) as follows

(3.16)

represents the "worst case" service rate of queue at the CPS when evaluating all the possible subsets of queues mapped onto active stages of the ancillary network at time .

Theorem 6 shows it exists a set of conditions which make the structure of the ancillary networks presented in this section respect Lemma 1 and, therefore, be an upper bounding network for the CPS.

Theorem 6. *Given a generic (i.e., not necessarily monotonic decreasing) CPS, and a ( ) feed-forward network associated to it, if , the following holds:*

(3.17)

then the feed-forward network upper bounds its associated CPS.

Proof: Also in the case of generic CPS, it is easy to show that the traffic arriving at stage

from previous stages, , is univocally determined by the state of the queues for fresh arrivals in the feed-forward network, . Furthermore, thanks to the policer, such traffic does not buffer at the GPS node. We can therefore indicate the service rate at time for fresh traffic at the -th GPS node as . Then can be written as

(3.18)

Eq. (3.18) ensures that the service rate of each queue for fresh traffic is always inferior to the one of the corresponding queues in the CPS, if the set of active queues in the CPS at time is a subset of the corresponding active queues in the network. Unfortunately, it is not said that is a subset of , the active queues in the ancillary networks. The only possible way for fresh arrivals at the network to be served at a faster rate than at the CPS is that at least one of the queues in is empty in the mapped queue of the network at time , i.e., it does not belong to the active queues in . If this is the case, the ancillary networks do not respect Lemma 1 and the networks described do not upper bound the generic CPS.

In the following, we prove that by contradiction. Considering that the arrivals are the same at the nodes of the network and at the corresponding queues of the CPS, we can assume that at , . Then, we assume by hypothesis that is the first moment at which a queue that is not empty at the CPS turns empty in the corresponding stage of the network. In other words, is the first moment at which does not hold. In order to satisfy the hypothesis , it exists at least a moment at which . For what previously said, this is possible just if at one of the queues for fresh traffic is empty in the ancillary network, while the corresponding queue at the CPS is not. The existence of contradicts the definition of , proving the theorem.

Exactly as for a monotonic decreasing CPS, it is easy to prove that the rate impairment traffic is univocally determined by the state of the queues for the fresh arrivals at the feed-forward network . In practice, even if Eq. (3.17)) depends on the time, it identifies a finite set of inequalities (one for each possible value of , i.e., conditions) that the ancillary network identified by ( ) has to respect in order to upper bound a generic CPS.

Note that the derivation of all results in Section 3.4 does not require the CPS to be monotonic decreasing. Therefore, using the new definitions of service rates and presented in this section, all the considerations made in Section 3.4 on the choice of the coefficients are still valid. Similarly, all sufficient conditions for stability and all analytical expressions of bounds derived for the monotonic decreasing case hold—with the newly defined and —also for a generic CPS.

Chapter 4

Modelling In-Band D2D Communications via D2D

In the last few years, many schemes aiming at enabling Device-to-Device (D2D) communications in cellular networks have been proposed. *In-band* schemes either allow D2D transmissions to occur over dedicated cellular resources (*in-band overlay* schemes [90]) or over the same resources used by legacy cellular users (*in-band underlay* schemes [91]). In this chapter we focus on the latter. Herein, we propose a queuing-theory based approach to performance characterization of D2D schemes, which captures the dependencies between interfering transmissions and achievable rates. Our approach is based on the Coupled Processor System (CPS) model [12] we presented in Chapter 3. Such model naturally applies to D2D systems, as it explicitly accounts for the achievable transmission rates when the correlation between the service rates of multiple queues is known. Our approach to CPS performance evaluation enables for the first time a fully analytical study of D2D communications.

The main contributions of this chapter are as follows. We present a novel analytical approach to the study of D2D schemes in LTE-like cellular networks, which applies to scenarios with one or more D2D pairs scheduled in the same Resource Block (RB) of a cellular user, or when little knowledge of traffic distribution is present. By applying our approach, we derive new sufficient conditions for stability of transmission queues in a D2D system, and we show how to evaluate the effects of D2D transmissions over cellular user performance. Moreover, we present a computationally feasible method for the determination of a proportionally fair allocation of resources, which allows trading the amount of fairness of the solution for computational complexity. Finally, we validate our results through simulations, assessing numerically the quality of the bounds and of the optimal allocations derived with our approach. Our results show the importance, in the performance study of a D2D system, of accurately modelling the mutual correlations in performance among users in a D2D cellular system.

4.1. System Model

We consider a base station (evolved Node B (eNB) in Long Term Evolution (LTE) terminology) belonging to an LTE access network, serving cellular users. Under the coverage area of the considered eNB, there are also D2D transmitter-receiver pairs. D2D pairs only have to transmit traffic to each other. The scheduling of the resources is centralized. The eNB, indeed, assigns resources both to cellular and D2D transmissions, even if D2D traffic does not have to traverse the cellular infrastructure. We assume D2D transmissions happen on the uplink channel. The main reason for this choice is that in uplink the eNB is the receiver of all cellular transmissions and has complete knowledge of the sensed interference. Therefore, when scheduling D2D transmissions, the eNB can exploit such knowledge to keep the overall interference under control [92]. As medium access technology, we assume that Orthogonal Frequency-Division Multiple Access (OFDMA) is used. Through OFDMA, the traffic of the users is spread over a time-frequency grid. Transmission time is split into slots of fixed duration, while multiple and independent sub-carriers are obtained over a wide channel bandwidth. A RB (which is a set of subcarriers over a single time slot duration) is the smallest resource that can be assigned to a particular user [1].

We consider a static scenario, in which D2D pairs do not change over time and in which users do not move. Nevertheless, the analysis may also be used in scenarios where users move. In those cases, indeed, the analysis would be valid for short intervals of time, where user channel quality can be considered static. The path loss model used is the log-distance path loss model. Specifically, path loss (in dB) between a transmitter and a receiver at a distance is given by — , where is the path loss at a reference distance , is the path loss exponent, and is a Gaussian random variable with standard deviation , modelling the effects of shadowing [93]. Finally, we model capacity through the Shannon formula. If a generic transmitter (either cellular or D2D) and its receiver are at a distance , the amount of bits transmitted per RB is:

—— (4.1)

where is the bandwidth of the RB, is the duration of a time slot, is the per-RB transmission power of , is the noise spectral density, is the inter-cell interference, while is the maximum amount of bits that can be transmitted in a RB when the best Modulation and Coding Scheme (MCS) is used. The summation of the interference at the denominator goes over all the active transmitters in the RB, both cellular and D2D. We assume that uplink power control is in place both at the cellular and at the D2D transmitters. Uplink power control achieves a particular

[1] The analysis holds also for Single Carrier - FDMA, where a set of RBs (and not one) is the minimum quantum of resource.

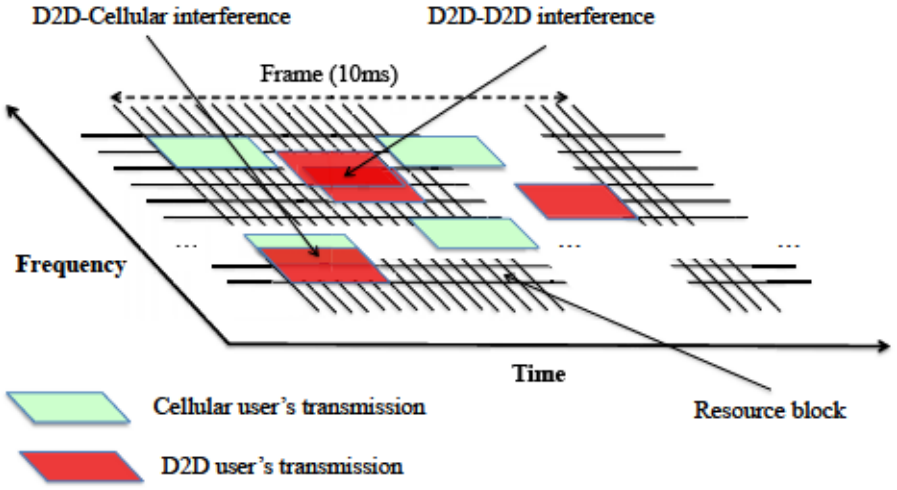


Figure 4.1: Use of resources in D2D in-band underlay systems.

per-RB SNR target, thus coping with path loss and shadowing at the receiver side. As a result, due to the proximity of transmitter and receiver in D2D, the transmission power used by D2D communications is typically small and generates little interference. Such interference becomes relevant in particularly dense scenarios, which are becoming real in today's networks. Dense scenarios are also the target of this chapter. Therefore, differently from many works available in the literature, we will not neglect the interference caused by D2D transmissions to other concurrent D2D and cellular transmissions.

4.1.1. A D2D In-band Underlay Scheme for LTE

In in-band underlay D2D transmission, a D2D pair can be scheduled by the eNB on the same RBs that are assigned to cellular transmitters, or to other D2D pairs. Fig. 4.1 represents an example of radio resource utilization of in-band underlay D2D communication.

The particular resource scheduling policy we consider is a variation over the one proposed in FlashLinQ [15]. FlashLinQ is a state-of-the-art PHY-MAC architecture for D2D that allows the scheduling of different transmitters (D2D or cellular) in the same time and frequency resource, through an OFDMA-like access selection mechanism. The scheduling of the transmitters is performed at RB level and is done taking into account the interference sensed by the receivers of the simultaneously scheduled transmissions. Nevertheless, FlashLinQ does not distinguish between cellular and D2D users. We believe instead that the UEs have to be considered as primary users of the cellular access network and that the D2D transmissions have to be scheduled without preventing satisfactory performance for cellular users. We introduce therefore a two-tier scheduling policy of the transmitters, where first cellular transmitters are scheduled independently and where, subsequently, the FlashLinQ scheduling policy is applied to the D2D transmissions. The overall modified version of FlashLinQ goes as follows:

- **Cellular User Equipments (UEs) transmitters**: The scheduling policy at the eNB is Equal Time [94], which is interference-unaware. [2] With Equal Time, all the UEs with traffic to send receive exactly the same quantity of RBs for transmission. In order to have a full understanding of the effect of the D2D transmissions on the cellular ones, we assume that cellular UEs are in saturation, i.e., they always have a packet ready to send. In practice, the eNB assigns RBs to each cellular user with the same probability.

- **D2D transmitters - FlashLinQ policy**: For each RB, all the D2D transmitters having at least one packet to send are considered. Those who are scheduled for transmission are all those whose interference on the cellular transmitter scheduled in the same RB is below a given limit, and who are able to achieve a given minimum Signal-to-Interference Ratio (SIR), considering the cellular UE scheduled in the particular RB under analysis and all the D2D transmitters already scheduled. This ensures that the impact of D2D transmissions on cellular ones is kept below a given threshold and, furthermore, that D2D users achieve a target minimum throughput.

The order in which D2D transmissions are considered by the above scheduling procedure determines the set of D2D scheduled transmissions in a given RB. Therefore, in order to maximize fairness among D2D transmitters, as in FlashLinQ, every time a scheduling decision has to be made, the order of candidate D2D transmitters is picked at random, following a uniform distribution.

4.2. A CPS Model for In-band D2D Transmissions

In this section the finding presented in Chapter 3 on the CPS queueing system will be used to tackle the analysis of in-band underlay D2D schemes.

A CPS is composed by a set of parallel queues, whose service rates at any time is determined by the set of active queues at that time. More formally, we denote the system *state* . is an array of binary variables, one for each queue of the CPS, which varies over time. At time the -th binary variable of is if the th queue have traffic to serve, and otherwise. The service rate of the -th queue of the CPS is univocally determined by , i.e., .

The D2D system can be mapped onto a queue-CPS (the number of queues exactly corresponds to the number of D2D transmitters under analysis). Each of the "coupled" queues models one of the D2D transmission queues, where the coupling arises from mutual interference with the other D2D pairs scheduled in the same RB. The CPS modeling of the D2D is sound because the the number of bits transmitted in a RB by the D2D pair can be univocally determined by the set of D2D transmitters active in the given RB.

[2] Any scheduling scheme, even interference aware, for which it is possible to determine the scheduling probability of a cellular user can also be used.

The number of bits transmitted in a RB by the D2D pair depends on the set of active D2D pairs, i.e., the system state , the cellular user scheduled in the same RB, and the particular order in which D2D users are listed by the scheduling algorithm. If we consider that cellular users are in saturation, then the average service rate of when represents the state of the pairs is:

— ——————— (4.2)

where and are respectively the number of active D2D pairs and the number of possible sorted lists of active D2D pairs when the system state is , is the number of available RBs per second. , computed applying Eq. (4.1), is the amount of bits per RB transmitted by when the cellular device is active and is the particular sorting used to schedule the active D2D pairs. takes into consideration the fact that the scheduling algorithm may not schedule D2D pair if the previously scheduled D2D pairs or the cellular user would incur an interference which is higher than the threshold interference set by FlashLinQ. In Eq. (4.2) we used the fact that all possible permutations of active D2D pairs are equally likely to be chosen, as well as the probability of having a cellular user scheduled in a RB is —, due to the Equal Time scheduler adopted.

Given Eq. (4.1), the approximation we make adopting the CPS model for the presented in-band D2D transmissions consists in considering the instantaneous service rate of the CPS queues as the average rate of the corresponding D2D transmitters in a given system state. Furthermore, while the cellular transmissions are scheduled sequentially, on a per-RB basis, and the change of system state can happen only when a new RB is scheduled, the equivalent CPS model works at the "fluid" limit, i.e., each queue would serve his traffic as it was infinitely divisible, and it could change state at any time . In Section 4.5 we validate numerically if those approximations are acceptable, and that the CPS models accurately the dynamics of the original D2D system.

4.3. Analytical Results

In Section 4.2 we showed how a system with cellular users in saturation and pairs can be described by a CPS. In the following we want to characterize the performance of the system. First of all we want to depict what is the set of achievable rates for the D2D pairs in the system. That is, we want to characterize the set of arrival rates at the D2D pairs which do not cause overflow to some of the backlogs in the system, i.e., the stability region of the system. Secondly, we evaluate the saturation throughput achieved by the cellular users when in-band D2D transmissions are allowed.

In the following, we assume that arrivals are upper bounded by leaky bucket arrival curves (Chapter 2, [14]). If is the D2D pair under analysis, is the leaky bucket rate and the burstiness. This assumption does not limit the applicability of our analysis. Indeed, in practical

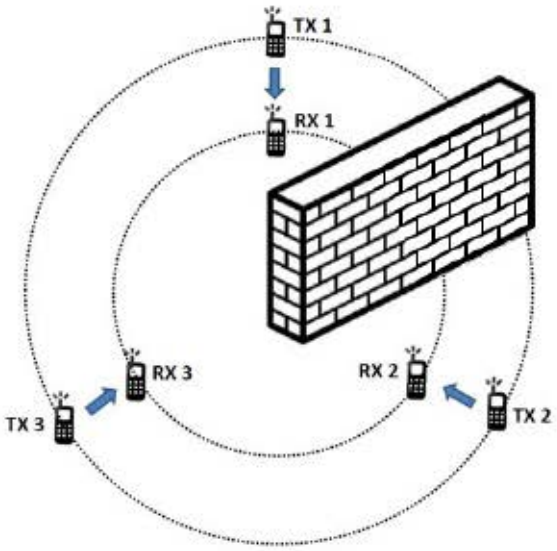


Figure 4.2: Non-monotonic scenario, with three transmitter-receiver pairs.

settings almost any source can be described by some leaky bucket arrival curve, possibly by means of some conservative assumptions on the statistics of the traffic (e.g., burstiness of the flow).

4.3.1. Stability Region of the In-Band D2D System

The underlying CPS system for an in-band D2D system adopting the FlashLinQ scheduler does not necessarily respect the monotonic property shown in Section 3.3. It is possible, indeed, that a particular D2D transmitter increments its achievable rate when the group of active D2D transmitters enlarges. To understand why this can happen, we present the following example. The setup is illustrated in Fig. 4.2 and it is composed by three different transmitter-receiver pairs. We assume a wall is present between pairs and , so that the reciprocal interference among the two pairs is heavily reduced. Due to the fact that the interference is very low, FlashLinQ always schedules simultaneously the two pairs on the same RB. The transmitter-receiver pair , instead, receives the other two pairs signals in line of sight and cannot be scheduled with any of the two. As a consequence, when the transmitter-receiver pairs and or and have traffic to send, they share equally, on average, the available RBs. If the remaining pair turns active, the transmitter-receiver pair is scheduled only one third of the times (when it occupies the first spot in the candidates ordering), while the other two transmitter-receiver pairs are both scheduled two third of the times. In such a setup therefore, turning active a transmitter-receiver pair improves the service rate of some of the queues in the system, giving rise to non-monotonic behaviors.

For a generic D2D system with pairs and cellular users we can obtain sufficient conditions for stability exploiting the analysis for generic CPSs presented in Section 3.5. For sufficient conditions for stability we mean a conservative bound for the actual stability region, described in terms of arrival rates. The following theorem introduces a set of inequalities ensuring that the

arrival leaky bucket descriptors , yield a stable behaviour.

Theorem 7 (Stability Region). *Consider an in-band underlay D2D system with pairs having traffic demands upper bounded by leaky bucket arrival curves with parameters , and the generic sorting of . For each sorting , the following inequalities yield a conservative estimate of the stability region:*

(4.3)

, and are the service rates of D2D pair computed as in Eq. (3.16) considering as set of active D2D transmitters, respectively, , and , while is the service rate of D2D pair when the set of active D2D transmitters is .

The union of all conservative estimates of the stability region, obtained under all possible permutations , lays within the stability region of the D2D system.

Eq. (4.3) is obtained by analysing the particular ancillary network introduced by our method in Section 3.5 with sorting (see Theorem 4). To obtain Eq. (4.3), the particular set of scalers used in the ancillary networks is the one we proved in Theorem 2 to satisfy the conditions imposed by Eq. (3.17). Hence, the ancillary network used yields upper bounding networks for the D2D system under analysis. Each sorting generates a different set of inequalities Eq. (4.3). Therefore, the union of the sufficient conditions for stability retrieved by the presented networks (one for each sorting) still yields a conservative estimation of the stability region of the in-band D2D communication system.

4.3.2. Saturation Throughput of Cellular Users

The model proposed in Section 4.2 for the activity of D2D pair can be used to obtain also a lower bound on the saturation throughput of each cellular user in the scenario, for a particular set of stable long term rates of D2D users . To achieve this goal, we build a new CPS model with queues, of which are in one-to-one mapping with the D2D transmitters and simulate the activity of the cellular users. In this modeling of the D2D system, we consider the traffic of the D2D pairs respecting the long term rates which proved to lead to a stable configuration of the system, i.e., respecting Theorem 7, while we let each of the cellular users always have some traffic ready to be sent.

Exactly as in Section 4.2, we study the CPS through a set of feed-forward networks such that, *at any time t*, the traffic of the cellular transmitters at the corresponding Generalized Processor Sharing (GPS) nodes is always served at most as fast as in the CPS. That is, at most as in the real D2D system. All the ancillary networks which prove stability for the D2D transmitters also provide a maximum rate for the cellular user which would ensure also its stability. If is the maximum arrival rate of queue for which we can prove stability through Theorem 4.3,

is also a valid lower bound for the saturation throughput of . Indeed, considering all the D2D transmitters traffic satisfied and the remaining cellular users as saturated, all the traffic entering the queue of would eventually exits its queue if was the actual arrival rate.

Given a particular set of long term rates for the D2D users, we build the feed-forward networks following the same mechanism presented in Section 4.2. In each of the feed-forward networks, the th GPS node now represents either a cellular or a D2D user, with capacity . In this particular case, is a sorting of . The expression of is computed as in Eq. (3.16) considering as the set of active transmitters mapped onto stages and all the cellular users mapped onto one of the stages . This slight modification also applies to the definition of and .

In order to accept as a bound for the saturation throughput of cellular user , has to be computed in an ancillary network which upper bounds the underlying CPS of the described D2D system and where *all* the D2D transmitters, with arrivals , are stable. When the particular ordering of is analyzed, a valid lower bound for the saturation throughput of cellular transmitter , mapped onto the th GPS node, is the maximum achievable rate computed from Eq. (4.3), i.e.:

(4.4)

Finally, the overall lower bound for the saturation throughput of is .

4.4. Proportionally Fair Optimization

By exploiting the knowledge of the conservative estimate of the stability region of the system, here we show how to achieve proportional fairness among D2D transmissions. Given a CPS with queues, each representing one of the D2D pairs in the scenario, we first formalize a proportional fairness throughput optimization problem that exploits the feed-forward networks introduced in Section 4.2, and then we present a heuristic which searches for the optimum by characterizing just a suitably small subset of networks.

4.4.1. Problem Formulation

We assume to have D2D transmitters, that their demand is given and that arrivals can be described through equivalent leaky bucket characterizations , . The goal of the optimization is to introduce leaky bucket shapers at the transmitters, having long term rates , in order to ensure stability at the transmission queues of the D2D users, maximize a weighted sum of the rates' logarithms, thus achieving *proportional fairness* of user's throughputs. The solution of the optimization problem and the distribution of the long term rates of the shapers to the D2D transmitters can be easily performed by the eNB where the D2D transmissions are taking place. Formally, we express the optimization problem as follows:

maximize

subject to: (4.5)

where the maximization is performed over any possible ordering . allows tuning the objective function in order to introduce different classes of users. The first constraint ensures that the long term rate we pick leads to a stable transmission queue for the generic transmitter D2D (directly from Theorem 7), while the second ensures that the long term rate we pick for the shapers is at most the one of the demand of the generic transmitter .

When dealing with a particular sorting , the presence of the function in the constraints leads to a piece-wise linear feasibility region. To deal with the problem, we use the Big-M transformation [95], so that the resulting optimization problem has a convex objective function and belongs to the mixed-binary programming family. We achieve the global optimum of the optimization through Branch and Bound ( ) method [96].

4.4.2. Derivation of a Heuristic

In order to reduce the computational complexity required by the optimization problem in Eq. (4.5), we stop the evaluation when the intermediate solution is at most at from the optimum (gives, at each step, a higher and a lower bound for the optimum). This parameter can then be tuned according to the desired trade-off between computational cost and performance of the heuristic.

The second and most important approximation we introduce is in the set of networks that we evaluate. Instead of choosing all the sortings , we try to identify the subset of feed-forward networks that most probably contains the stable set of long term rates that maximize our proportional fair problem. We limit to those networks the evaluation of the optimization problem in Eq. (4.5).

To choose such subset we use the following reasoning. Due to the cascade structure of the feed-forward network, for each CPS queue, the earlier is the stage to which the corresponding GPS queue belongs, the larger the set of queues whose coupling with the considered one is modelled more accurately, i.e., through feed-forward traffic rather than via a penalty on the service rate which holds for any time .

Obviously, the higher the weight of the D2D pair in the proposed optimization, the higher is the contribution to the utility function of the system. Therefore, the objective of the heuristic we propose is to maximize the achieved throughput of the D2D pairs associated to the largest weights.

Table 4.1: Simulation Setup

LTE Carrier UL Bandwidth B Subframe Duration (64-QAM with coding rate 0.93) Min. SIR packet length	 mW

Accordingly, the small subset of networks we decide to evaluate is initialized by probabilistically mapping D2D transmitters onto GPS nodes sequentially, starting from the earliest stages of the feed-forward network, with probabilities proportional to the weights of D2D transmitters in the objective function of the optimization problem in Eq. (4.5). We repeat the probabilistic mapping process to populate a small set of starting points for the heuristic. For each starting point network, we identify the set of "neighbour" networks each obtained by swapping two adjacent GPS nodes. Within this set of neighbours, the network with the higher proportional fairness for the system is taken as the new reference network, and the evaluation of neighbour feed-forward networks is repeated until a local maximum for the proportional fairness is found. Obviously, the trade-off among computation and accuracy of the solution is given by the number of starting points we choose.

4.5. Numerical Evaluation

We are now ready to validate the accuracy of our CPS-based model for in-band underlay D2D communications. First, in a small scenario, we simulate a real D2D system to evaluate the tightness of bounds on stability region, delay and backlog achieved as described in Section 4.2. Then we characterize the D2D system performance in terms of stability region and of saturation throughput of the cellular users. We pick a small scenario for graphical purpose, although all the results we show can be easily replicated for a larger number of devices. Finally, we show the advantage achieved exploiting our analysis when a proportional fair assignment of the resources is the goal. For the whole set of proposed simulations, Table 4.1 summarizes the values of the parameters we use. We also use power control, aiming to achieve a per-RB Signal-to-Noise Ratio (SNR) of .

4.6. Model Validation

In what follows we validate the bounds achieved by our approach on a simple D2D scenario. The scenario we consider consists in transmitter-receiver pairs residing on the vertices of two concentric d-gons, of radius respectively and , where is the distance between each

transmitter and its receiver. For the particular results shown in the following, , while . In this first evaluation, no cellular user is present in the scenario.

As we said in Section 4.1, we adopt a log-distance path loss model, with being the loss exponent. We assume devices do not move, and channel characteristics do not change over time. In such a system, interference at a receiver (and therefore the throughput achieved by each transmission queue) does not vary. As a result, interference is determined only by the set of active transmitters in the system, making the CPS model a good fit for performance analysis of such wireless network. Each of the queues of the CPS models a transmitter-receiver pair. The number of bits that a transmitter can send out in an RB is determined by the particular ordering used to analyze the transmitter-receiver pairs in the FlashLinQ scheduler. Only the D2D transmitters which respect the SIR threshold imposed by the scheduler are eventually allowed to transmit on a given RB. Eq. (4.1) is used to compute the particular throughput achieved by each D2D transmitter. In the particular example we analyze, interference is computed each RB over the set of scheduled pairs.

Independently from the number of D2D transmitters present in the system, we assume an obstacle is present between pairs and , so that the path loss exponent between the two pairs grows significantly (and the reciprocal interference among the two pairs reduces) as a function of the signal attenuation induced by the obstacle. Again, Fig. 4.2 illustrates the case .

Note that this setup generates non-monotonic behavior and that is why it has been selected in this section. With FlashLinQ, a transmitter-receiver pair is scheduled on an RB only if it does not cause high interference to the neighboring receivers. For the particular chosen topology, where several D2D pairs are in each other proximity, if and are low (e.g., or), the interference signal at the receivers is such that only a transmitter-receiver pair can be scheduled at once. Undoubtedly, the CPS model for such wireless scenario is monotonic decreasing, since introducing a new transmitter-receiver pair to the scenario reduces the time of transmission of all the previously present pairs. Non-monotonic behaviors, which we will analyze in the next section, arise instead when assumes larger values. Table 4.2 shows the service rates of each D2D pair in the system, when , for different values of and for each possible system state. Table 4.2 also shows the computation of applying the Eq. (3.16), valid for generic CPSs.

4.6.1. Modeling D2D via Monotonic Decreasing CPS Setup

In this section, we evaluate the performance of the system when the underlying CPS is monotonic. In particular, we assume that and . Therefore, we assume the scenario to be completely symmetric. When , we evaluated the sufficient conditions for stability of the system when the arrivals are constrained by a leaky bucket or when the arrivals were characterized stochastically. To achieve such result, we used Theorem 4 and Theorem 3, respectively. In order to obtain the largest possible characterization of the sufficient condition for stability, we did not limit the choice of the scalers to the ones suggested by Theorem 2, but

Table 4.2: Service rates Fig. 4.2 if and .

TX	System State	Serv. Rate Mb/s ( )	Serv. Rate Mb/s ( )	Serv. Rate Mb/s ( )	Mb/s ( )

we randomly selected scaler sets around the values of the scaler family . As we said previously, each scaler set and each sorting leads to a completely different ancillary network. The final definition of the bounds for the stability region was obtained as the union of the stability region of each ancillary network.

Fig. 4.3 shows two sections of the boundaries of such regions, derived by keeping constant the arrival rate for one of the three transmitters, respectively at and . For comparison, we also plotted the stability region obtainable assuming a full saturated system, as well as the stochastic stability region derived by an alternative CPS approach in the state of the art [44]. These plots show that our approach allows increasing sensibly the set of input rates for which the system is known to be stable, with respect to the well-known stability region obtainable assuming coupling effects are always those of a saturated system. Such improvement is more marked the farther the system is from the saturation point. Indeed, in a monotonic CPS, a decrease in arrival rate at a node frees up some system capacity for other flows, by decreasing the effect of coupling on service rates at other nodes in the system. Hence these plots show that our approach is accurate enough to retain such a property of the CPS, which accounts for the better performance over saturated models such as [44]. Fig. 4.3 (a) shows that, as expected, our sufficient conditions (for both stochastic and deterministic stability) are exactly equal to existing results for two nodes CPS, as discussed in Section 3.4.1. As those conditions are tight, they define a region (shadowed in the plots) for which the system is known to be unstable. Note that [44] performs well only when the demands of *all* of the queues is under the saturation point. Indeed by [44] any point for which the demands of two or more nodes are above saturation is to be considered unstable. Such approach is therefore not efficient if the number of nodes in the system increases, because the probability of having a stable set of demands where the demands of two or more queues are above the demand that can be satisfied under saturation increases.

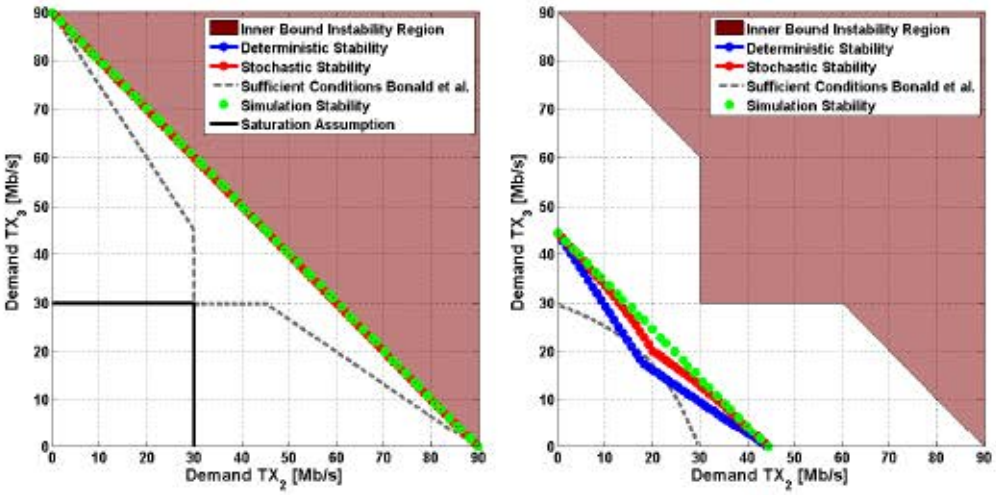
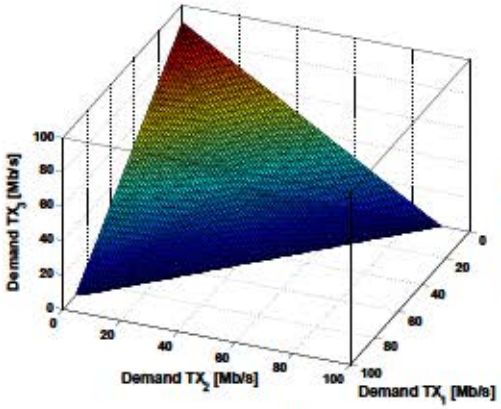


Figure 4.3: Stochastic and deterministic stability regions, for a demand at Tx 1 of 0 Mb/s (a) and Mb/s (b).

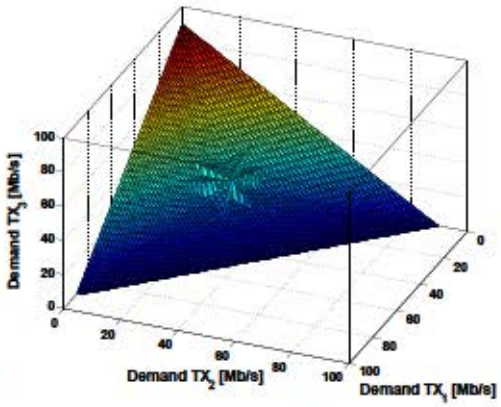
In order to have an indication of the tightness of our sufficient conditions for stability, we have run some simulations of the presented D2D communication system. We used an event-driven simulator in MatLab, which serves at any time the queues associated to each transmitter-receiver pair with a rate that is univocally determined by the set of active queues at , computed as in Eq. (4.2). We assumed Poisson arrivals and we observed the evolution over time of queue length at nodes for arrival rates in the interval . is the maximum rate achievable by a D2D transmitter if it receives all the resources in the LTE spectrum for himself.

For each combination of arrival rates, we have adopted the following crude heuristic in order to individuate system configurations which could be associated to instability. We have divided the simulation time (one hour) into five intervals, discarding the first in order to reduce the effect of the initial transient. For each of the remaining intervals, we have computed the average queue length over the time interval. A set of rates has been considered likely to lead to instability when, for at least one node, such average queue length increased steadily over the four intervals. As we can see from Fig. 4.3, the border of the region individuated by such heuristic are very close to the borders of the stochastic stability region given by our sufficient conditions. Such empirical estimations suggest that our conditions are reasonably tight, and that the set of arrival rates for which queues do not explode in practice is not much larger than the one individuated by our bounds.

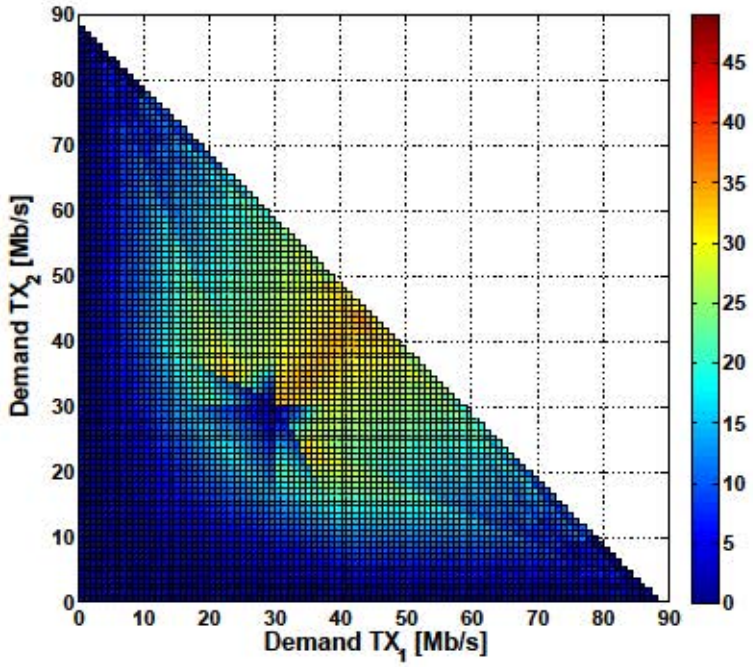
In Fig. 4.4 we show the surface plot of the stochastic and the deterministic stability regions, both derived by optimizing the bounds in we used Theorem 4 and Theorem 3 over the scaling values. As such bounds are based on different assumptions on input traffic, and as they relate to different notions of stability, in general they are not comparable. We can get an idea of their relative performance by considering the special case of leaky bucket constrained traffic sources,



(a) Stochastic stability region



(b) Deterministic stability region



(c) % Difference between stochastic and deterministic maximum rates for a transmitter, in function of the arrival rates for the other two.

Figure 4.4: Deterministic and stochastic stability regions, for the three node case.

with leaky bucket rate equal to their mean traffic arrival rate. To a system with such input traffic characterization, we can apply both Theorem 3 for its stochastic stability, and Theorem 4 for its deterministic stability. Fig. 4.4 (but also Fig. 4.3) show that our stochastic stability conditions are looser than those for deterministic stability, particularly around the saturation point. The mean relative difference is , with the bounds for deterministic stability being always tighter than those for stochastic stability.

In order to have an idea of how the performance of our bounds changes with respect to the number of nodes of the CPS, we have evaluated numerically the stochastic stability region for

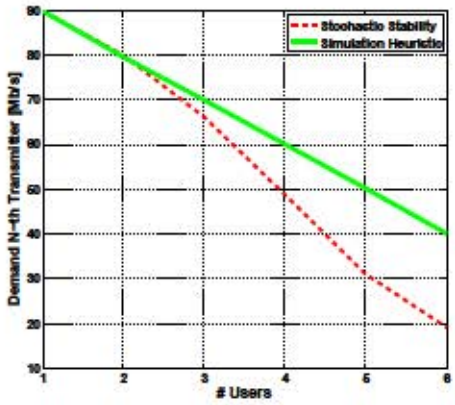


Figure 4.5: Maximum demand of the -th user for which the system is stochastically stable, as a function of . The demand of all other users is Mb/s.

. For every value of , we have set the demand of transmitters to , and we have computed the maximum demand of the -th user for which the system is stochastically stable, according to our bounds, as well as the one resulting from our heuristic on simulations.

Fig. 4.5 shows that the difference between our bounds and simulation results increases with . Indeed, our analysis is based on a worst-case model of the behavior of each node. As our analysis is based on the feed-forward structure of the upper bounding networks, the effects of the worst-case approach cumulate at each stage, bringing to results which are more conservative as (and therefore the number of stages of the feed-forward networks) increases.

Finally, we have assessed numerically the results on mean delay and backlog at CPS queues of Section 3.4.1, optimizing them over the scaling values satisfying Theorem 1, for the case of . In order to observe how mean delay and backlog evolve when the system approaches instability, we have chosen fixed (random) ratios between the demands of the three transmitters, with each ratio individuating a specific linear trajectory in the three-dimensional space of user demands. Along such trajectories we have evaluated the mean delay and backlog from Section 3.4.1, as well as those resulting from simulations, using our previously described heuristic to determine when to stop simulating in the trajectory.

In Fig. 4.6 we show the results for the trajectory with the greatest gap between simulation and analytical values. Simulation values are averaged over simulation runs. The demands are expressed in function of , with , , and . The plots show that simulation results are very close to analytical values, with the latter being always larger than the former. This is expected, given that our results are based on a worst-case approach. We also see that, as the trajectory does not pass from the saturation point, the increase of the mean delay and backlog with load is faster at one node (transmitter in the considered trajectory) than at the others. Moreover, at this queue the sharp increase in mean delay and backlog from simulations takes place very close to the border of the stability region given by Theorem 3. All this suggests that the sufficient

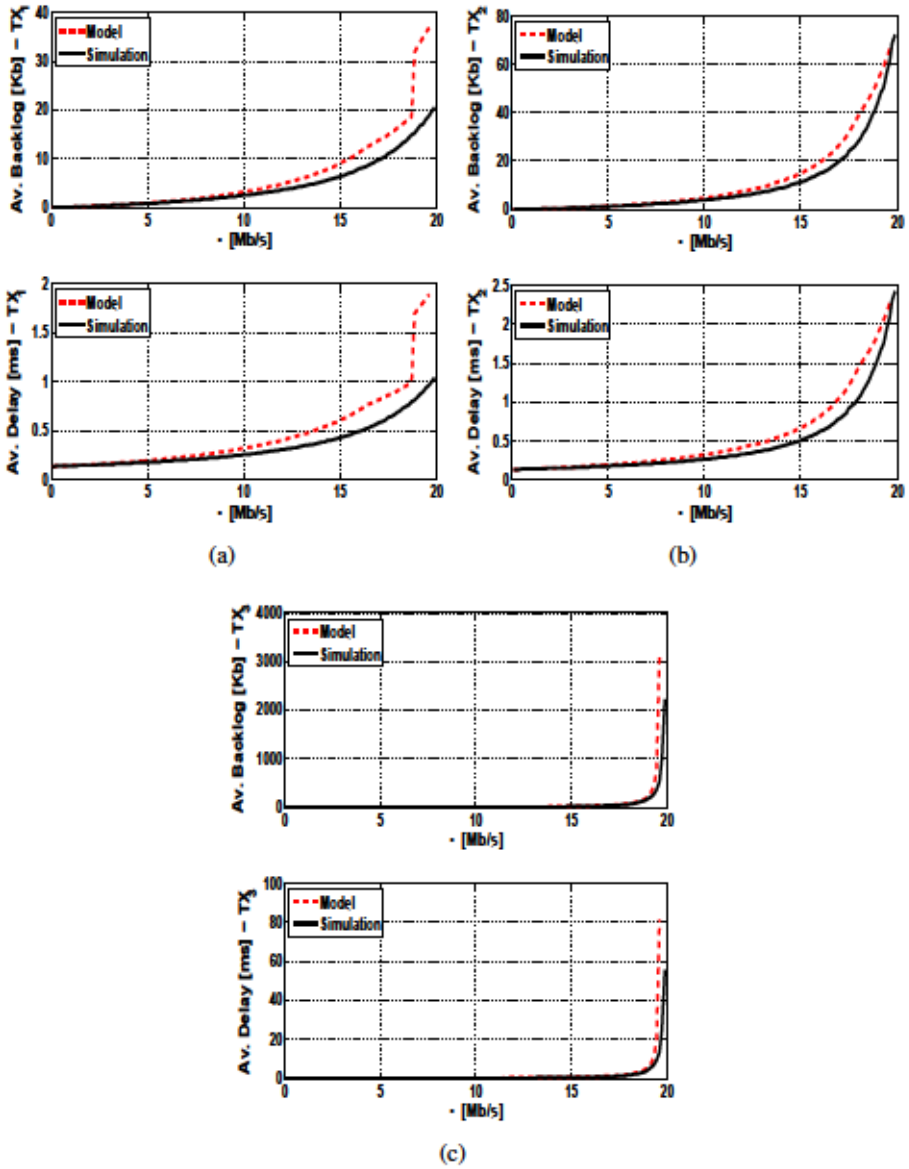


Figure 4.6: Mean delay and backlog (analytical vs simulation) when $\lambda_1 = \lambda$; $\lambda_2 = 1.5\lambda$; $\lambda_3 = 2\lambda$: (a) Tx 1, (b) Tx 2, (c) Tx 3.

conditions for stochastic stability and the bounds resulting from our approach are reasonably tight.

4.6.2. Modeling D2D via Non-monotonic CPS

We now analyze a wireless setup leading to a non-monotonic CPS. We consider the path loss exponent of the obstacle among pairs 1 and 2 being $\eta_w = 4$. As it is easy to see from Table 4.2, pair 1 and pair 2 reach 60 Mb/s when also pair 3 has traffic to transmit. Nevertheless, when or pair 1 or pair 2 do not have traffic to transmit, the service rate of the other pairs decreases to 45 Mb/s.

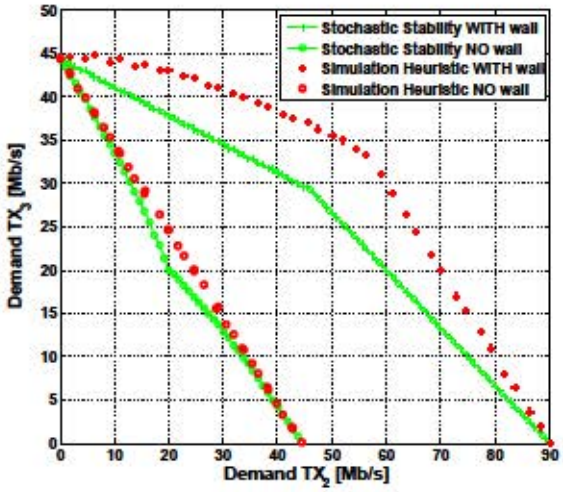


Figure 4.7: Non-monotonic vs. monotonic stability region.

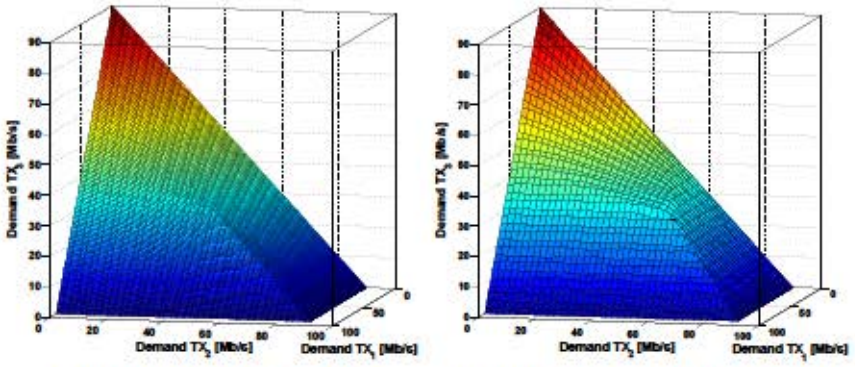
Using Theorem 6, we are able to derive the set of feed-forward networks needed to analyze the presented non-monotonic D2D scenario and from Theorem 3, we computed the inner boundary of the stochastic stability region. In order to have an idea of the tightness of the achieved inner bounds of the stability region, we also run a set of simulations for estimating the stability region according to the same heuristic as in Section 4.6.1.

Fig. 4.7 shows the same section of the stability region as presented in Fig. 4.3(b), i.e., fixing the demand of transmitter-receiver pair to . Even if the analytical boundary underestimates the service rates of the transmitter-receiver pairs and when all queues are fully backlogged, the average difference among the analytical boundary and the simulated set of stable demands is , i.e, the . Such a percentage is high only because a small difference in among the analytical and the simulated values is very high in percentage when the achievable throughput of the third pair is very limited.

Fig. 4.7 also allows comparing the stability region achieved by the system with and without the presence of the obstacle in the scenario. Clearly, the reduction of interference among the pairs and enlarges the region of demands that can be considered stable both from the analytical and from the simulations point of view.

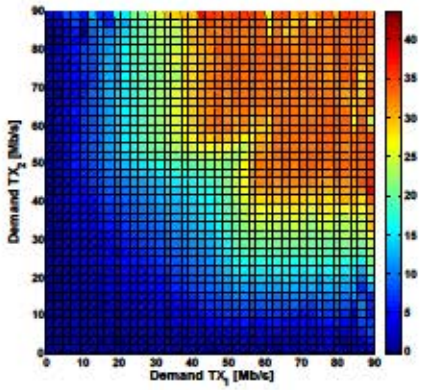
Finally, Fig. 4.8 represents the full characterization of the inner bound of the stochastic stability region and the simulated one. In particular, Fig. 4.8(c) shows the difference (in) among the two. The analytical maximum stable demand of the transmitter-receiver pair is always smaller than the one computed by simulation. This confirms that our analysis leads to a conservative estimation of the real stability region of the system also in the case of a non-monotonic CPS. Furthermore, the highest difference among the analytical stability region and the simulated one is . The difference is mainly present when the demands of the transmitters and , i.e., the transmitters whose service rate are not monotonic decreasing, are higher. Again, in those conditions, a very small difference in achievable throughput is instead very high on percentage.

Furthermore, the analysis presented constitutes the first available approximation result which is valid also for non-monotonic CPSs, and the fact that the difference among the model and the simulations is quite small is very promising.



(a) Stochastic stability region

(b) Heuristic Simulation stability region



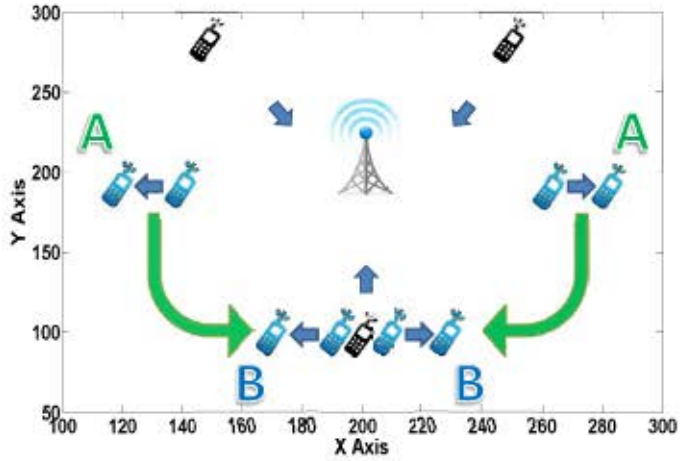
(c) % Difference between stochastic and simulated maximum rates for a transmitter, in function of the arrival rates for the other two.

Figure 4.8: Stochastic and simulated stability regions, for the non-monotonic three node case.

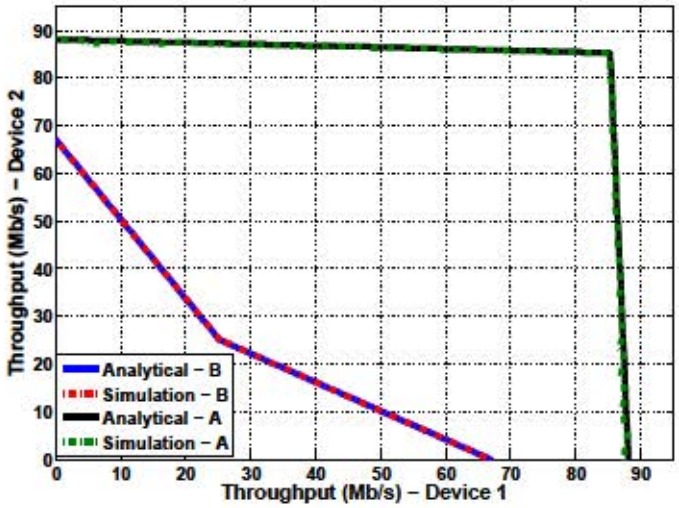
4.6.3. A Scenario with Cellular Users: Stability Region and Saturation Throughput

We start now introducing some cellular users within the evaluated scenario. We pick a scenario where cellular transmitters and D2D pairs are present. The position of cellular transmitters (UE UE UE) and eNB is fixed as in Fig. 4.9. The distance among cellular transmitters and eNB is exactly the same for the UEs ( ). In different simulations, the D2D transmitters are moved symmetrically from position to position , although their distance from the eNB is also fixed to . By moving the D2D pairs closer and closer, we want to evaluate the impact of coupling among transmissions.

To evaluate the impact of distance for D2D transmissions, Fig. 4.9 shows our conservative



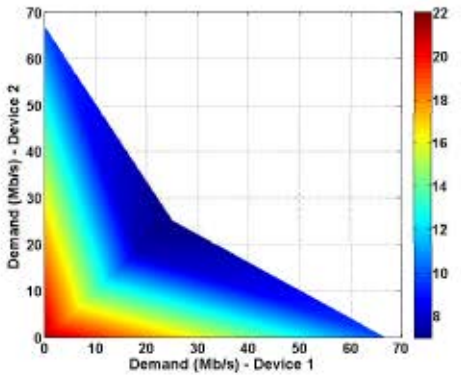
(a) Position of the devices.



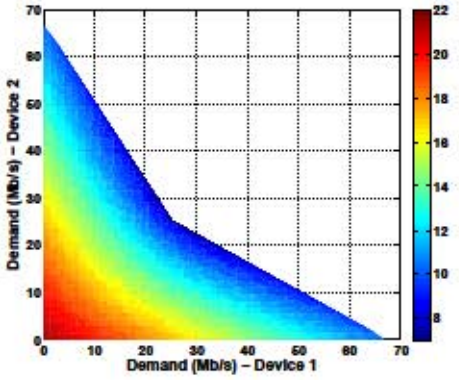
(b) Stability Region.

Figure 4.9: Scenario under analysis.

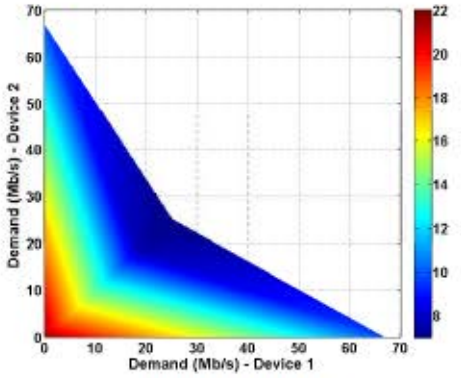
estimate of the stability region for D2D pairs and the rates achieved in the simulator at the two extreme positions we studied (position , maximum distance; position , minimum distance). The bounds on the stability region follow the results in Section 3.5, while the region depicted by the simulations has been obtained using the same approach as in Section 4.6.1. As expected, the stability region in the two cases is extremely different. When widely distant, the D2D pairs do not influence each other and they substantially increase the spectral efficiency of the cell (both are able to transmit almost in each RB using the maximum rate). When close, the D2D transmissions heavily impact on each other, and a small increase of the demand of one of the two causes a sensible reduction of the throughput achieved by the other. Regarding the quality of the stability region achieved analytically, the difference of the areas shown in the picture is and , when the D2D transmitters are in position and , respectively. We also evaluated



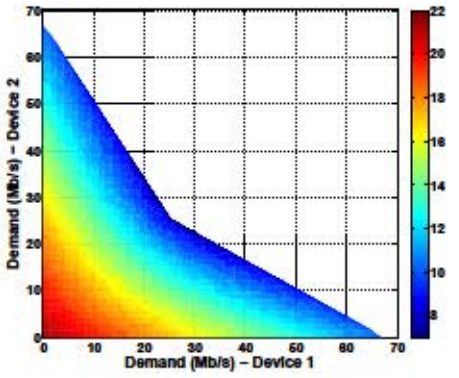
(a) UE TX1 (Model).



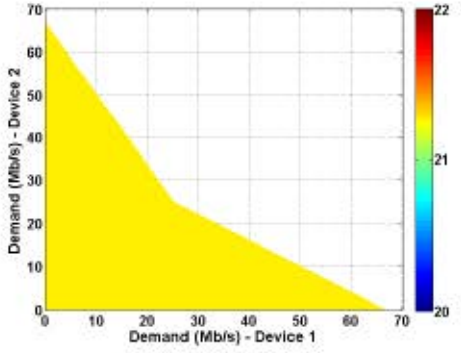
(b) UE TX1 (Simulation).



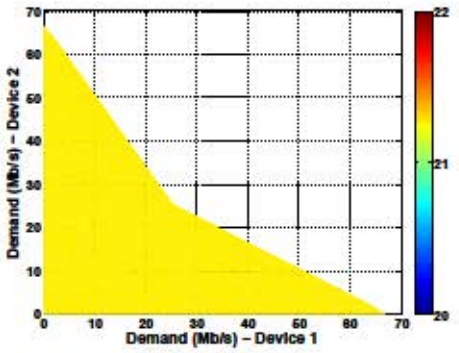
(c) UE TX2 (Model).



(d) UE TX2 (Simulation).



(e) UE TX3 (Model).

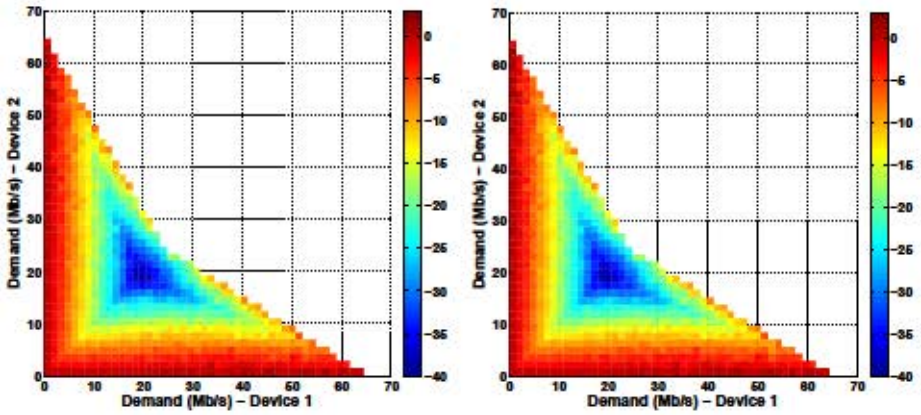


(f) UE TX3 (Simulation).

Figure 4.10: Maximum Achievable Throughput, Cellular UEs. Analysis vs. Simulations.

randomly generated topologies with 3 cellular users and 2 D2D users, and the biggest difference observed was as little as .

To dig into the results, Fig. 4.10 represents the effect of the D2D pairs over the cellular transmitters when the D2D transmitters are in the most coupled position, i.e., in position in Fig. 4.9. The figure shows both the conservative estimate achieved by means of the introduced model, and the saturation throughput achieved by simulation. Inspecting the results, it is clear that



(a) UE TX1. (b) UE TX2.

Figure 4.11: Cellular saturation throughput differences between analysis and simulations (percentage w.r.t. simulations). The analysis is conservative, although large underestimates only occur when saturation throughputs are relatively low.

UE TX3 achieves always the same throughput, independently from the D2D demands. Indeed, due to the proximity of the D2D receivers to UE TX 3, the scheduling policy does not allow D2D pairs to transmit so to avoid poor rates. It is possible to evaluate from the particular chosen case the benefit that D2D transmission brings in throughput terms. Specifically, for each demand set of D2D transmitters, it is easy to compute the throughput reduction suffered by cellular users. However, even if the cellular transmitters loose just little achievable throughput in total, the sum of D2D throughputs is potentially very high (greater than 60 Mb/s). Hence, in-band underlay D2D ensures efficient utilization of resources.

Fig. 4.11 quantifies the difference among the conservative analytical estimation of the saturation throughput and the throughput achieved via simulations by cellular users (see Fig. 4.10). In the case under analysis, the worst per-user underestimation goes as far as , although the average underestimation is as low as and large underestimates only occur for low saturation throughputs. All other evaluated cases presented similar values for the underestimation of the saturation throughput of the cellular users, and the worst per-user average underestimation we observed was .

4.6.4. Optimization

Now, we show the results of the optimization problem we presented in Section 4.4. In the following, we stopped the approach at most far from the optimal solution. From different experiments, represents a good trade-off among precision and complexity. First of all we present how close the heuristic performs if compared against a brute force approach, i.e., an approach that solves the optimization problem over all the possible upper bounding networks, and with a gap to the optimal solution for the algorithm equal to .

In order to show the scalability of the approach we propose, here we choose larger scenarios.

Table 4.3: Heuristic vs. Brute force: Utility

D2D transmitters	Heuristic		Brute force	
	Mean	Conf. Int.	Mean	Conf. Int.
3	2.53	2.38-2.69	2.53	2.38-2.69
4	2.57	2.45-2.71	2.58	2.45-2.72
5	2.34	2.21-2.47	2.35	2.22-2.47
6	2.21	2.02-2.41	2.23	2.04-2.43

Table 4.4: Heuristic vs. Brute Force: Complexity

D2D transmitters	Mean # of Networks (Heuristic)	Confidence Interval	Networks Available
3	3.44	3.36-3.52	6
4	4.60	4.36-4.84	24
5	9.56	8.95-10.18	120
6	11.70	10.49-12.92	720
8	22.70	20.52-24.88	40320
D2D transmitters	Mean # of Branches (Heuristic)	Confidence Interval	Branches Available
3	2.00	2.00 - 2.00	2
4	1.99	1.90 - 2.07	4
5	2.74	2.56 - 2.93	8
6	2.85	2.55 - 3.15	16
8	4.82	4.10 - 5.53	64

In Table 4.3 we present the results obtained when cellular transmitters and , , and D2D transmitters are present. The choice of the number of transmitters is such that we can solve the optimization problem also with the brute force approach. The position of the devices, the demand of each of the D2D transmitters and each weight in the utility computation is picked at random for any of the optimization problems performed. Table 4.3 shows that the presented heuristic performs as good as the brute force approach in most of the cases, although it does not need to explore the entire set of ancillary networks.

In order to evaluate the complexity of the heuristic proposed, Table 4.4 shows the number of networks evaluated by the heuristic and the complexity of each of the optimization problems solved (one for each network). Such complexity is evaluated in terms of different branches the algorithm requires before reaching an intermediate solution that is at most far from the optimum. In this case we also evaluate a set of larger optimization problems, with D2D transmitters. In particular, Table 4.4 shows that the complexity of the solution proposed by the heuristic presented in Section 4.4.2 remains computationally feasible, even though the complexity of the brute force approach grows fast.

In order to show the gain that the knowledge of the conservative estimate of the stability region brings, we also simulate all the scenarios for which the optimization was performed. We simulate, for each of the cases, the operation of the in-band underlay D2D system, considering or not the presence of shapers at the D2D transmitters set as the output of the optimization problem we solved. In both cases we compute the throughput achieved from the D2D transmitters, and

Table 4.5: Optimization vs. Saturation: Utility

D2D transmitters	Mean Sat.	Conf. Interval	Mean Opt.	Conf. Interval
3	2.32	2.11-2.54	2.53	2.37-2.68
4	2.28	2.06-2.50	2.57	2.44-2.70
5	1.98	1.80-2.17	2.34	2.21-2.47
6	1.74	1.39-2.09	2.21	2.01-2.41
8	1.51	1.25-1.77	2.05	1.91-2.19

then we compare the corresponding log-utility (see the optimization problem in (4.5)).

As it easy to see from Table 4.5, even if we achieve just a conservative estimate of the whole stability region, the shapers improve sensibly the value of the utility achieved. In particular, when the scenarios get larger, the utility improves up to on average and in the best case. Please note that the simulations performed when the shapers are present reach almost the same utility values that the optimization problems were giving in output.

Chapter 5

A CPS Model for 802.11 Out-band D2D Communications

Wireless mesh networks based on IEEE802.11 are nowadays an inexpensive, well widespread solution to easily, effectively and wirelessly connect entire cities. Thanks to such pervasiveness, they are poised to play a central role in many "Internet of Thing" application scenarios, with very diverse Quality of Service (QoS) requirements. Many of those application, as for example Machine to Machine communications, have been also proposed to be integrated within the next generation of cellular networks for an easy and always-on user experience [97]. Out-band IEEE802.11 Device-to-Device (D2D) communications may therefore become a common practice in cellular networks and, as we show in Chapter 6 and in the second part of this thesis, become a crucial mechanism to improve their performance.

The wide deployment and the potentials of out-band IEEE802.11 D2D communications makes it crucial to develop models for analytical performance study of such networks. Empirical studies in complex environments hardly give clear indications of the general properties of wireless mesh networks. In the state of the art, performance analysis of the 802.11 Carrier Sense Multiple Access - Collision Avoidance (CSMA-CA) mechanism has traditionally focused on saturated traffic assumptions [10]. Many of the available results for non-saturated conditions do not capture the effects of traffic dynamics on system performance.

In this chapter we propose a different approach, which is based on a Coupled Processor System (CPS) modeling of the IEEE802.11 MAC behavior. Contrarily to the state of the art approaches, our modeling allows capturing the dependencies between user achievable rates and traffic dynamics, due to sharing of the wireless transmission medium (mediated by the CSMA-CA mechanisms). For the analysis of the underlying CPS modeling of the IEEE802.11 MAC, we use the approach proposed in Chapter 3.

The main contributions of this chapter are as follows.

- We present an analytical approach to the study of IEEE802.11 networks under non-saturation conditions. By applying our approach, we derive sufficient conditions for stabil-

ity of transmission queues.

- We present a computationally feasible method for the determination of a proportionally fair allocation of resources in ad hoc networks, which allows trading the amount of fairness of the solution for computational complexity.

- We validate our results through simulations, assessing the quality of the bounds and of the optimal allocations derived with our approach.

The chapter is organized as follows. Section 5.1 introduces the system model, and the main assumptions underlying our analysis. In Section 5.2 we introduce our method based on the CPS model, and we derive sufficient conditions for stability of our network. Section 5.3 presents a heuristic for the computation of a proportionally fair operating point for the system. Finally, in Section 5.4 we assess numerically our results.

5.1. System Model

We consider a scenario with hosts communicating via the IEEE protocol in ad hoc mode. We assume all hosts are in range of each other. We assume hosts do not suffer for additional interference on the IEEE bandwidth and that they do not move. Such an assumption is in line with many application scenarios, as for example, Machine to Machine communications.

We do not make any assumption on the traffic generated by applications at each host. However, we assume that such traffic is passed through a *leaky bucket controller* [14], before being sent through the wireless connection. This device forces its output to be constrained by a *leaky bucket arrival curve*, with parameters (burstiness) and (rate). We mentioned the properties of a leaky bucket arrival curve in Section 2.2. In few words, if is the cumulative arrival rate at the output of the controller, , , [14]. At the controller, all arrivals from the application layer which are not conformant to such arrival curve are buffered. Introducing such controller allows to tune the amount of traffic buffered at the transmission queues of the hosts and, possibly, to avoid some pathological conditions for the system. A well-known problem of IEEE , for instance, is that few hosts with poor channel quality jeopardize the medium, at the expense of all other hosts. In these cases, constraining the traffic sent to the network by each host might help achieving a better allocation of system resources. Leaky bucket controller could be implemented at the application layer through communication among nodes, adjusting the operating point of the system in order to optimize a given utility function (e.g., in order to achieve fairness). Moreover, assuming some form of constraint on sources allows dividing hosts into classes, with different service levels for each class.

5.2. A CPS model for Out-band D2D Communications

In this section we present how to model IEEE as a CPS. We use the model to study the performance when the system is not saturated, i.e., when the hosts not always have traffic to transmit. In this particular case, CPS is able to capture the coupling among hosts due to the IEEE MAC protocol.

5.2.1. The CPS Model

As mentioned in Section 3, a CPS is composed by a set of parallel queues (i.e., queues which do not exchange traffic among them) served by work conserving schedulers, and whose service rates at any time is completely determined by the set of active queues at that time. We define as the *state* of the system at a given time the array of binary variables . At time , the -th binary variable of is if the -th queue have traffic to serve, otherwise. Then at time the service rate of the -th queue is , i.e., it is only function of the state of the system at time .

In what follows, we model our -node out-band D2D network as a -queue CPS. The application of the CPS modeling to out-band D2D, even though is intuitive, has never been presented before this thesis. The state of such CPS is given by the set of active D2D transmitters at time . In modeling our out-band D2D system as a CPS, we assume that for each D2D transmitter at time , the service rate is completely determined by the state of the system. Since the system is not working in saturation, the set of active queues generally changes over time. Nevertheless, when the set of active queues is the same at different intervals of time, the performance of the active hosts are roughly the same (apart from a short transient period). Therefore, we assume that the throughput achieved by the hosts in a given system state is equal to the throughput they would achieve when the subset of active hosts is in saturation. For each set of active hosts there exists a different saturation throughput, and the system oscillates among the different system states depending on the set of hosts with traffic to send. For each set of active hosts, we characterize the saturation throughput as in [10]. The saturation throughput is valid for IEEE , but it can be easily modified to be valid also for IEEE . If is the set of active transmitters at time , the instantaneous service rate at the -th active transmitter is given by:

(5.1)

Here, the denominator represents the average duration of a slot of the IEEE MAC. is the average slot duration when a transmission is successful, is the average slot duration when a collision happens on the channel, is the probability that a transmission occurring on the channel is successful, which is given by the probability that exactly one station transmits on the channel. is the probability that there is at least one transmission in the considered slot time. Finally, is the duration of an empty slot time. The numerator instead represents the average

number of bits transmitted by host in a time slot and is the average packet payload size. All these parameters can be computed directly from the parameters of the CSMA-CA protocol, and they are strictly dependent on the set of active transmitters . For the expressions of each parameter, please refer to [10].

The main approximation of using the CPS modeling for out-band D2D communications is that the CPS model works at the "fluid" limit, i.e., each queue serves its traffic as it were infinitely divisible, while the wireless system follows the MAC protocol of IEEE which, for example, lets a single host transmit at once. In Section 5.4 we assess the validity of such assumption, showing numerically that these approximations model accurately the performance of out-band D2D communications.

5.2.2. Sufficient Conditions for Stability of an Out-band D2D System

Eq. (5.1) allows characterizing the underlying CPS system of the IEEE network we want to analyze. For the particular case of IEEE , whenever a new host turns active, the average throughput of the hosts which were previously active decreases. For this reason, out-band D2D communications can be modeled as a monotonic decreasing CPS, and the results achieved in Section 3.4 on the conservative estimation of the stability region still hold.

In what follows we recall the main result theorem of Section 3.4 which we exploit in the following section. Let us assume to have a characterization of the arrivals of the hosts in out-band D2D system given by the set of of leaky bucket parameters . When the particular ordering of the D2D hosts is used, the out-band D2D system is stable if:

(5.2)

and are defined as in Section 3.3.2, while the schedulers can assume any value respecting Theorem 1. The result is directly obtained applying Theorem 4 to the out-band D2D system.

5.3. Derivation of the Optimal Operating Point

Choosing the leaky bucket parameters of traffic sources in our networks allows tuning the operating point of the system. For example, leaky bucket parameters can be chosen in order to maximize some utility function as, for example, a fairness utility function. IEEE already implements some form of fairness through the use of frame aggregation. Theoretically, when frame aggregation is in place, the same amount of resource is assigned to all users (in terms of airtime) and proportional fairness is achieved. Unfortunately, frame aggregation implementations deviate largely from the expected behavior, depending on the vendor [98].

The following technique, instead, achieves fairness through rate limiting techniques at the application layer, overcoming vendor-dependent differences.

5.3.1. Problem Formulation

In the present section, the utility function that we choose to optimize is a weighted fairness function, which is one possible way of balancing some notion of fairness among users with, for instance, different classes of service. Its expression is:

— (5.3)

where is the minimum bit rate for an acceptable performance for the application. The feasible set of leaky bucket rates over which to optimize such utility is given by the set of inequalities in Eq. (5.2), as they define the set of rates for which the out-band D2D communication system can be considered stable.

The feasible operating points which maximize the weighted fairness are therefore the solutions of the following optimization problem, computed over the set of ancillary networks we can gather analyzing the underlying CPS.

maximize —

subject to: (5.4)

———————

where the constraints derive from Eq. (5.2). This optimization problem is similar to the one we proposed for in-band D2D communications. Therefore, also this optimization presents two challenges: the presence of the min function in the constraints, which leads to a non-convex feasibility region; and the complexity of the problem, which scales factorially with the number of hosts.

5.3.2. Heuristic Approach

In order to practically solve the above problem, we propose a heuristic approach that it is very similar to the one proposed in Section 4.4.2. We describe the heuristic also in this section for the sake of completeness.

The heuristic is composed of two parts. The first part aims at reducing the problem to a tractable problem that can be solved with standard tools. Specifically, since the presence of the function in the constraints of the above problem leads to a non-convex feasibility region, we use the so called *big-M transformation* [95]. Instead of that constraint, the method builds two

constraints in which we add a binary variable multiplying a large constant value . Whenever the binary variable is equal to one, the large constant makes the constraint useless because all the feasible sets of leaky bucket rates satisfy it, while when the binary variable is zero the constraint is active. By choosing a value for the binary variables we select a part of the feasibility region. Therefore, the two terms of the min in Eq. (5.4) are never *active* at the same time.

The problem obtained in this way belongs to the mixed-binary programming family. We solve it by means of Branch and Bound ( ) method [96]. For each , we stop the branch-and-bound evaluation when the intermediate solution is at most at away from the optimum. Tuning allows achieving different tradeoffs between computational cost and optimality of the solution.

The second part of our heuristic aims at reducing the number of auxiliary networks over which to search for the optimum. It is based on running a set of greedy searches from a set of starting points, each of which has been derived as follows. To each node of the CPS, we associate the quantity —. Then, starting from the first stage of the auxiliary network, we assign a node of the CPS to each stage with a probability proportional to this quantity. The idea underlying such algorithm for the choice of the starting points is that nodes with higher weights in the utility function need to be modeled more accurately than the others. This is achieved by assigning those nodes to the last stages of the auxiliary network. Indeed, due to the structure of the ancillary network, the lower the stage a node belongs to, the larger the set of nodes whose coupling with the considered one is modeled through accurate rescaled traffic rather than via a conservative penalty on the service rate, holding for any time , and therefore independent on traffic patterns at interfering nodes.

We describe now the elementary step of the search. From a network , we consider the set of networks obtained by a swap of two contiguous nodes in and we perform the optimization in Eq. (5.4) for the resulting ancillary networks. If the log utility value of is lower than the max log utility among all these networks, the network with the highest log-utility value among all the is selected. Otherwise, the search stops. The largest of all local maxima computed from all starting points is the final output of our heuristic. By changing the number of starting points we can achieve different trade-offs between computational cost and optimality of the solution.

5.4. Numerical Evaluation

In this section we assess numerically our results. First, we evaluate the fitting of the proposed CPS model for a IEEE network. In particular, we evaluate the impact of assuming the saturation throughputs in Eq. (5.1), state by state, as the service rates of the CPS equivalent model. Then, we evaluate the performance of our heuristic, for proportional fairness. The operating points derived with our approach are compared to those obtainable in saturated condition, i.e., assuming all queues are always active, as a function of the number of hosts in the system. Moreover, whenever feasible, i.e., for settings with only few nodes, we have compared our results

Table 5.1: Setup of the Wireless Scenario

Available Channel Speeds (Mbit/s)	1;2;5.5;6;9;11;12 18;24;36;48;54
CWmin	16
Backoff stages	5
Preamble + PHY header( )	20
SIFS ( )	16
ACK Time ( )	24
DIFS ( )	34
Slot time ( )	9
MAC header(bits)	224
Chunk size (bits)	15000

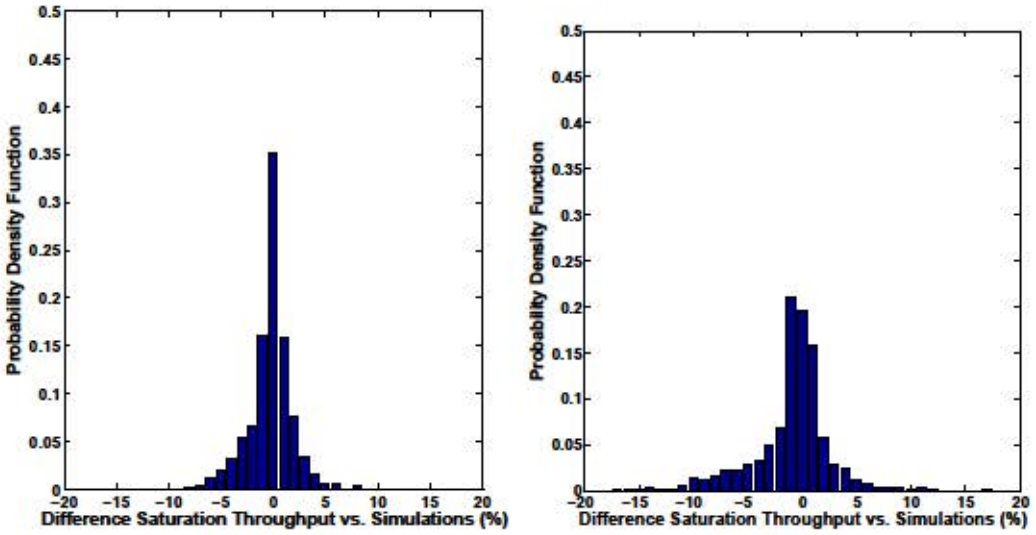
with those obtainable by solving the optimization problem in Eq. (5.4) through brute force over all the ancillary networks as presented in Section 5.2.1.

The parameters used in the considered scenarios are presented in Table 5.1. We have chosen the standard for hosts communication. The values of the parameters are derived from [99].

5.4.1. CPS Model Validation

In this section, we evaluate the fitting of the CPS model for communications. The evaluation is performed via simulation. We implemented in MATLAB the MAC of and we selected at random for each of the transmitters a fixed channel quality among the ones of Table 5.1. For different scenarios and given the selected channel qualities selected, we pick at random a stable set of arrival rates for the transmitters. The decision of taking stable arrival rates ensures that the set of active transmitters present in the wireless scenario, i.e., the system state, changes over time. Then, for each system state and for each transmitter, we evaluate the difference among the average throughput achieved during simulations and the service rate used in the CPS model, i.e., Eq. (5.1). Fig. 5.1 presents the Probability Density Function (PDF) of the differences for scenarios having and nodes. Similar results have been obtained also for different set-ups.

In both cases, the service rates used in the CPS modelling are close to the ones computed during simulations. In particular, in the and the of the cases, respectively, the absolute value of the difference among the service rates computed as in Eq. (5.1) and the corresponding ones achieved through simulations is less than the . Even if the system assumes a given state just for a short interval of time, i.e., even if the subset of active transmitters change quite quickly, the above result shows that the saturation throughput of Eq. (5.1) is reached fairly



(a) 3 Nodes. (b) 7 Nodes.

Figure 5.1: Probability density function, difference Simulations vs. CPS service rates.

soon in almost all the cases. Therefore, we consider as negligible the impact of the approximations introduced during the modelling of the system as a CPS.

5.4.2. Proportional Fairness Optimization Results

In this section we evaluate the results achieved through the optimization problem Eq. (5.4) in a large set of scenarios. In each setting, we solve Eq. (5.4) exploiting our heuristic, computing for a given number of users the set of leaky bucket rates which maximize the log-utility function.

In the following, the weights of the utility function, defined in Section 5.3.1, are uniformly distributed in . Instead of considering a specific propagation model, and modelling its impact on the achievable channel rates, for each user we assigned channel rate randomly, assuming rates to be uniformly distributed among the set of available rates in Table 5.1. We have set the parameter for the to . Furthermore, for a given number of hosts , we considered a number of starting points for the heuristic, i.e., a number of ancillary networks, which scales with . Indeed, as grows, it increases also the solution space, as well as it does the space of possible values of the weights of the utility function, and of the channel rates. Empirically, and for the number of users considered in our evaluations, we have found that scaling the number of starting points as — brought acceptable results in terms of output of the optimization and of computational complexity.

Overall, for each value of we considered a total number of instances of our setting (i.e., a particular choice of starting points, set of weights and set of channel rates) sufficient to get a confidence interval within the of the value of the average utility achieved. In any case, we never used less than 100 instances. A first objective of our numerical evaluation has been to assess the performance of the proposed heuristic, which has been introduced as a

Table 5.2: Average Log-Utility: Heuristic vs. Exhaustive Search

Opt. technique	3 Nodes	4 Nodes	5 Nodes	6 Nodes
Ex. Search	3.4690	3.0688	2.8034	2.3933
Heuristic	3.4690	3.0590	2.7879	2.3630
Difference	0	-0.32%	-0.55%	-1.27%
Max Difference	0	-6.56%	-10.64%	-15.88%

computationally feasible approach to the problem of maximizing the (weighted) proportional fairness in the allocation of leaky bucket rates among IEEE hosts. More specifically, we have tried to give an idea of how far, on average, the solutions of our heuristic are from the optimal values, in order to evaluate the impact of the approximations on which the heuristic is based. To this end, for scenarios with a small number of nodes (for which an exhaustive search still brings to an acceptable computational complexity), we have compared the average log-utility derived through our heuristic with the one derived through exhaustive search over all the possible upper bounding networks and choosing . From Table 5.2 we observe that, in the considered scenarios, the solutions from our heuristic bring a utility which is on average very close to the optimal values derived through exhaustive search. This suggests that the approximations on which our heuristic is based have an overall low impact on the optimality of the operating point derived.

In Fig. 5.2 we compare the average log-utility, together with the confidence interval, from our heuristic and the one obtainable in the saturated scenario. Fig. 5.2 also contains the median of the utility in the same cases. We can see how in all cases the average log-utility derived by optimizing (through our heuristic) over the set of operating points which are stable according to our method is always at least larger than the one derived by assuming the system in saturation. Moreover, we see that the relative improvement (in terms of utility) brought by our heuristic over the utility achieved under saturation assumption grows with the size of the scenario. The larger is the number of the stations in the system, indeed, the higher is the rate of contentions and, consequently, the inefficiency of the MAC under saturation assumptions. As we can see from the evaluation of the median in Fig. 5.2, the difference between the optimal values of the utility function derived with these two methods is relevant also in distribution, and even larger than the one for the average.

Applying leaky bucket controllers at the stations reduces the number of always contending stations and therefore improves, as this result confirms, the efficiency of the wireless channel. Nevertheless, it is interesting to note that for both methods the optimal value of log-utility decreases when the number of nodes increases. Therefore, also when the heuristic is applied, the devices face an increase of the inefficiency of the MAC, even if at a smaller scale.

In order to have a better idea of the difference between the operating points resulting from the heuristic and from the saturation assumption, we have compared them on the basis of the total average weighted throughput, in order to take into account the relative contribution of each host

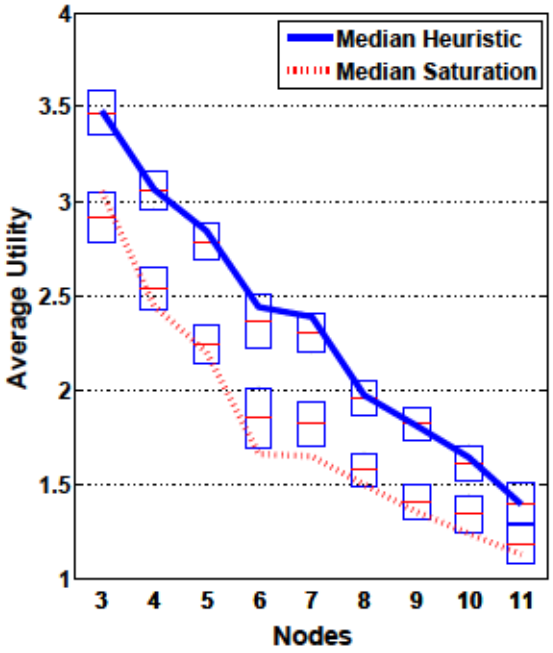


Figure 5.2: Average and Median Log-utility. Heuristic vs. Saturation approx.

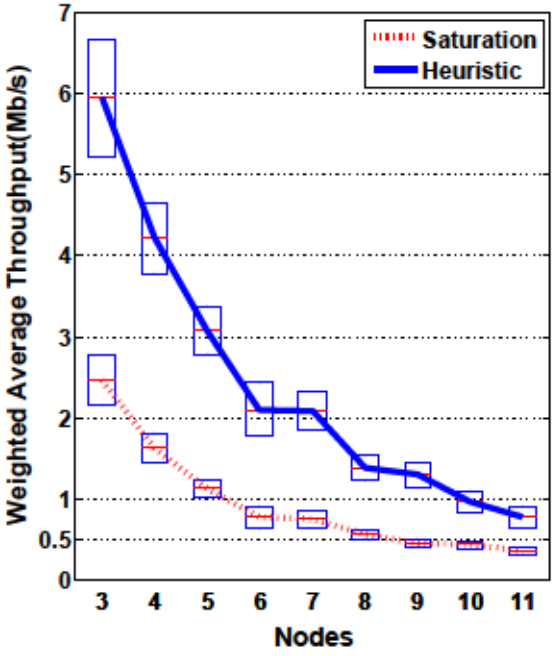


Figure 5.3: Weighted Av. Throughput. Heuristic vs. Saturation Condition.

to the utility of the system. That is, the weights in these sums are the same as those adopted in the utility function. The results are shown in Fig. 5.3, where the cases under analysis are exactly the same used in Fig. 5.2. We see how our heuristic brings the system to an operating point for which the total average throughput is at least higher than the total average throughput imposing

Table 5.3: Complexity of the Heuristic

	Average Number Network - Heuristic	Maximum Number Network ( )
3 Nodes	6	6
4 Nodes	17.15	24
5 Nodes	41.04	120
6 Nodes	59.22	720
11 Nodes	409.05	
	Av. Per-Network branches - Heu.	Av. Per-Network branches - Exhaustive Sear.
3 Nodes	4	4
4 Nodes	6.49	8.32
5 Nodes	9.47	12.87
6 Nodes	11.89	17.80
11 Nodes	29.11	-

saturation throughput as rate limits. This shows how our heuristic derives system operating points that, besides maximizing the utility function, bring to a much more efficient utilization of network resources.

In order to understand the feasibility of the proposed approach, we analyzed the complexity of the heuristic we propose versus the exhaustive search of Table 5.2. In Table 5.3 we present first the average number of networks the two methods analyze in order to get the final log-utility of the system. In case of the heuristic, we count the total number of upper bounding networks analyzed, considering all the — starting points. We also present the number of optimization problems which the method solves for each network. In the worst case, the number of branches visited by the for nodes is . Results are shown up to the point at which the comparison is computationally feasible (6 nodes), and for the largest scenario we analyzed through our heuristic (11 nodes). We can see how the heuristic requires a considerably inferior number of evaluations (both in terms of upper bounding networks, both in terms of branches), with a very limited impact on the optimality of the value of the log-utility derived (as seen in Table 5.2). Please note that the different starting points of the heuristic are completely independent from each other. Therefore, the computational time can be reduced sensibly if the heuristic is evaluated in parallel. Finally we evaluate how the results change varying the number of starting points and the value of . Also here, we evaluate all the scenarios discussed formerly when and nodes were present in the network. The results presented in Table 5.4 use the setting having and — starting points as a benchmark. Even though the utility remains almost

Table 5.4: Dependence of Results on and on Number of Starting Points (SP) heuristic

	4 Nodes	7 Nodes
, , —		
, , —		
, , —		
, ,		
Weight. Av. Th., , SP = —		
Weight. Av. Th., , SP = —		
Weight. Av. Th., , SP = —		
Weight. Av. Th., , SP =		

the same in every configuration of the heuristic, the weighted average throughput changes sensibly. The most affecting parameter of the heuristic is clearly the number of starting points used. Increasing the number of starting points the weighted throughput increases accordingly. Unfortunately, the number of upper bounding networks increases of the and with and nodes, respectively, leading to a computationally expensive resolution of the proposed optimization. However, considering as acceptable the complexity of the proposed heuristic, reducing the number of starting points results in an unnecessary lost in performance, that can become tricky when the number of nodes increases. The same reasoning applies when we evaluate the choice of : sometimes the difference of performance is too small to justify the decreasing of the value of . For instance, the increase of branches analyzed when reaches the if compared to , while the accuracy of performance estimates remains similar. The presented heuristic represents, therefore, a good trade-off among complexity and performance.

Chapter 6

Energy Efficiency in Mixed Access Networks

Multi-technology access networks with heterogeneous protocols and resource allocation schemes are sprouting. On the one hand, telco operators already implement -based hotspots alongside the traditional cellular infrastructure [17]. On the other hand, the standardization of Device-to-Device (D2D)-assisted cellular networks is now becoming reality [4] and future cellular networks (e.g., 5G) are envisioned to support D2D communications.

In this context, D2D-assisted cellular access networks might allow bundling user traffic in groups and implement dynamic relay, without requiring additional infrastructure deployments. If so, they would allow improving the service towards destinations experiencing poor or no channel quality. Contrarily to what we analyzed in Chapter 4 and Chapter 5 where D2D was analyzed only as a mean of communication among users, in this section we analyze an access network where *D2D can be seen as an alternative network access technology.*

In the following, we focus on access networks in which D2D complements the presence of standard Access Points (APs) and evolved Node Bs (eNBs) via out-band connectivity. We refer to such networks as *mixed access networks.* Devices providing network access are referred as Points of Access (PoAs).

We start our work by deriving a novel comprehensive model for throughput and terminal power consumption estimation in mixed access networks which accounts for the intertwined nature of cellular and resource allocation when D2D comes into play. In fact, while the throughput achieved by users in the cellular network by D2D relay nodes affects the downlink load of D2D over channels, it is also true that the performance of mobile terminals affect the quantity of traffic that can be offloaded, and therefore cellular users. The model measures also energy efficiency because battery lifetime becomes of paramount importance in mixed access networks. While acting occasionally as D2D PoA, since they have to receive and relay also the traffic of other User Equipments (UEs), mobile devices experience a higher power consumption.

Our model includes an innovative approach to the characterization of cliques under any traffic load condition, using a simple yet effective description of as a system with a variable number of fully backlogged terminals. The approach represents an approximation of the analysis presented in Chapter 5 and uses standard tools like Bianchi's analysis [10]. With our model, we compute the *airtime* used by terminals for communicating to any kind of PoAs, even when experiencing heterogeneous channel qualities and under variable load conditions. The airtime characterization is novel and key in our proposal, since it is needed to compute throughput and power consumption of the terminals.

Building on our model, we formulate a new network-controlled access selection problem that aims to maximize the energy efficiency achieved at terminal side. As a major point, in the proposed formulation, not only uplink and downlink are both simultaneously considered, but also they are jointly optimized in terms of cellular and resource allocation. Since the resulting problem is NP-Hard, we design a novel and efficient heuristic for access selection in mixed access networks. We validate our model and the heuristic via packet-level simulations and show that, thanks to D2D-enabled relay, the access selection mechanism we propose achieves up to higher efficiency and much fairer throughput distributions than standard access selection procedures. Notably, our results scale well with the density of terminals in mixed access networks.

The reminder of this chapter is structured as follows. Section 6.1 describes the components of network-controlled mixed access networks. Section 6.2 presents our innovative and comprehensive model. In Section 6.3 we formulate the energy efficient access selection problem and propose an on-line heuristic. In Section 6.4 we validate the model and present numerical results.

6.1. Mixed Access Networks with SDN

In this section, we present the components of a mixed access network in which eNBs, APs and D2D connectivity are concurrently deployed. Mixed access networks represent one of the most likely solution to many issues arising in the next generation of cellular access networks. In order to provide higher broadband connectivity to mobile users, cellular networks are steering towards a densification of eNBs. However, jointly with the densification of device populations—e.g., due to the emerging paradigms of Internet of Things and machine-type communications—this tendency creates higher interference and increases the probability that users far from an eNB experience low-quality channel conditions. Therefore, integrating the classical cellular bandwidth with the spectrum of APs or implementing D2D connectivity allows to sensibly improve the average spectral efficiency and deliver far better user experience than with cellular-only resources.

In a mixed access network, users directly connected to an eNB can act as relay for uplink and downlink transmissions of other users, using D2D. In any case, users can attach to a single PoA, i.e., an AP, an eNB or a D2D relay. We denote by Mobile Node (MN) the user attached to an AP, and by UE the user attached to an eNB, whereas for relays and users attached to relays we use WiFi-Direct terminology, i.e., Group Owner (GO) and Group Member (GM), respectively. Each

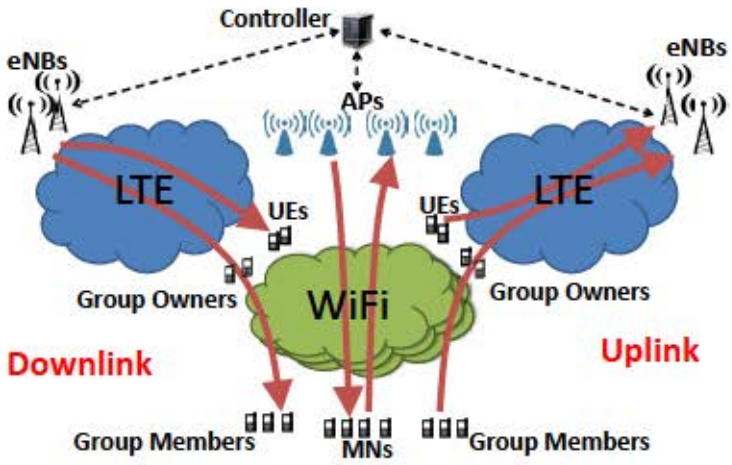


Figure 6.1: SDN-controlled mixed access network.

UE can become GO and form and manage a single WiFi-Direct group to receive and transmit uplink and downlink traffic of all the users (GMs) accessing the network through that GO. D2D transmissions in a WiFi-Direct group use frequencies and therefore do not interfere with cellular communications. Nevertheless, different groups, as well as different APs, can be assigned to the same channel and contend for the same resources.

Unfortunately, and cellular resources are limited and it is possible that GOs are not able to relay all the received traffic in uplink and/or downlink. In order to avoid wasting valuable resources, we assume that in a mixed access network, operators are the ones steering the resource allocation, both in the cellular network and in the channels. In the considered mixed access network, resource allocation relies on SDN, and a controller has the goal of collecting relevant pieces of information from the PoAs to enforce traffic policies dynamically. Fig. 6.1 illustrates the above-mentioned network elements.

Specifically, exploiting the knowledge of channel qualities of all the terminals in the system, as well as their uplink and downlink demand, the controller applies the analytical approach we present in Section 6.2 to estimate the throughput each user can achieve under the current PoA association. Thanks to the computation of the achievable throughput of the users, the controller enforces rate limiting at the sources and ensures that no resource is actually misused. The SDN controller is also the entity that manages the access selection decisions. In the particular mixed-access network we enviosion, access selection is performed with the objective of increasing battery lifetime at the terminal side. The algorithm we propose to perform access selection is presented in Section 6.3.

6.2. System Model

We start by proposing a novel model for the estimation of throughput and power consumption at terminal side which is valid in the proposed mixed access network. We first model cellular and

throughput and power consumption in isolation, i.e., under given load levels. Then we model the impact of the coupling between cellular and due to D2D relay.

6.2.1. Analytical Framework

Network. We denote by the number of eNBs in the access network, whereas represents the set of eNBs. The set is the set of users attached to eNB . The number of GMs attached to the WiFi-Direct group managed by is denoted by . If is not a GO, . The uplink demand of is . Such demand includes the uplink traffic of plus the uplink traffic of the GMs attached to it. The downlink demand is instead indicated as , which includes the demand of and of all its GMs.

We denote by the number of APs/GOs using the same channel , whereas represents the set of APs/GOs on the same channel. We assume that all the APs/GOs using the same channel are in radio range with each other. If not needed, we omit the index when analyzing a given clique. The set is the set of users attached to a particular AP/GO . The traffic demands of are indicated as and in uplink and downlink, respectively. Therefore, the notation used for users is similar to the one used for cellular users, although the exact meaning will be clear from the context.

Cellular transmissions. We assume that power control is operated only in the uplink channel, and that uplink and downlink occur on different frequencies. Following [100], we assume that the terminals set their transmission power in order to achieve a particular Signal-to-Noise Ratio (SNR) target, thus coping at the receiver with path loss and shadowing. Furthermore, we assume that all the eNBs use the same cellular bandwidth, i.e., frequency reuse 1 is used. We assume that the cellular transmissions adopt Orthogonal Frequency-Division Multiple Access (OFDMA) to allocate resources to the users. Nowadays, this is common practice in 3GPP for 4G networks, and next generation of cellular networks (5G) will continue on the same path. We also assume that eNBs are always active in downlink (e.g., due to intense download activity), whereas the uplink might be not saturated (e.g., due to bottlenecks between the GMs and their GOs).

Cellular scheduling. We assume that eNBs operate independently from each other and that use a Vertical First Scheduler (VFS) [101], i.e., a subframe is allocated, when possible, entirely to the same user. Due to the fact that it reduces the airtime by design, VFS achieves high energy saving. In the VFS framework, users are scheduled with a Weighted Round Robin (WRR) policy.

Scheduling of GOs. Since GOs haul the traffic of other users, we assume that their weight in WRR is proportional to the number of GMs connected plus one (the GO itself).

D2D and operation. In , we assume that no power control is used. Furthermore, we assume that APs or GOs serve their users in Round Robin (RR) order. In this chapter we also assume that is used.

cliques. We assume that, due to the use of different channels or due to distance, we have independent (non-interfering) sets. Nevertheless, GOs and GMs, as well as MNs and APs belonging to the same set compete for the same resources. For this reason,

even if they do not have to transmit to each other, they form a clique.

Demands and arrival processes. We assume demands and arrival processes are stationary and independent, but not identically distributed for each user and for uplink and downlink. Arrivals are packets with the same average size for all cases.

Uplink and downlink coupling. Contrarily to what assumed in the majority of the work on access networks, here we analyze the uplink and downlink jointly, accounting for the fact that they are coupled. For example, if the downlink throughput achieved by a GO over the cellular link changes, we consider the change in the downlink relay, which will affect the uplink traffic from GMs, and, consequently, the uplink load over the cellular link of the GO.

Controller role. The controller is able to estimate analytically (as we show in the following) what is the maximum amount of traffic that can be actually relayed by the GOs, both over the cellular and the channel they have access to. Therefore, the controller is also the entity that enforces rate limiting at the sources accordingly to the maximum achievable uplink and downlink throughput of everyone, avoiding waste of resources. If rate limiting is not in place, a GO can be assigned cellular resources for relaying traffic which may not be relayed in the corresponding channel (e.g., because of channel saturation). The same holds also for traffic, which may not find available resources in the cellular bandwidth. Finally, the controller is also the entity that makes access selection decisions for the new terminals entering the system.

It is worth noticing that different assumptions on the scheduling, resource allocation, packet size or on the network control operation can be easily accommodated in our model.

6.2.2. Analysis of Cellular Operation

The throughput of cellular UEs is strictly related to the number of resources assigned to a UE (proportional to its activity time, the airtime) and the number of bits that a UE is able to transmit for each resource (bit efficiency). Similarly, the power consumption under VFS scheduling is proportional to the airtime of the terminals [101]. Therefore, we develop an analytical method to compute bit efficiency and airtime.

6.2.2.1. Uplink

Let us first consider uplink transmissions to the eNB. Let be the transmission power used by UE , which is the result of the uplink power control mechanism in place and it is set to ——— , where is the target SNR that each user tries to achieve in uplink transmissions [100], is the noise power, is the path loss in the transmissions from to eNB , while is the maximum power transmission allowed.

We now show how to calculate the average uplink bit efficiency achieved by each uplink transmitter, that is, the average number of bits that each UE can transmit in a OFDM symbol. The average level of interference sensed by each eNB is the power received from the users attached to the other eNBs. Thus, using the Shannon formula over the bandwidth of a symbol and over

its duration , the resulting bit efficiency is:

(6.1)

where is the uplink airtime of an interfering user in a different cell, and the second term in the argument of the function is the SINR.

With VFS, the airtime used in uplink is the portion of symbols needed to serve the throughput , out of a total of the symbols transmitted by the eNB per second over the entire uplink bandwidth. Since the number of symbols per second required is , the airtime is:

(6.2)

The above formulas show that there is a complex relation between airtime, bit efficiency and throughput. Indeed, the bit efficiency and the airtime achieved by a particular user depend on the bit efficiency and the airtime of the interfering users, which in turn depend on the bit efficiency and the airtime of the user under analysis.

Denoting by the unsatisfied demand (in terms of symbols/s), the airtime is also expressed as:

(6.3)

in which can be computed iteratively. We propose Algorithm 1 to solve the problem without approximations. The algorithm, which is a recursive fixed point algorithm, estimates iteratively the uplink airtime of each user in the scenario and its achieved average bit efficiency. The algorithm starts by assigning users airtime proportionally to the relay group size . Then, given the resource allocation at the previous step, Algorithm 1 estimates the bit efficiency, given the bit efficiency just computed, an updated version of the resource allocation. In the algorithm, the available symbols/s are distributed to the users respecting the weights of the WRR scheduler up to satisfying the demand. Unused resources are then assigned to the users whose demands are higher than their fair share.

Upon convergence, or after a maximum number of iterations, Algorithm 1 returns the airtime allocation that corresponds to the served demand (computed as in line 20 of the algorithm's pseudocode). The throughput can be then computed by inverting (6.2).

Finally, the uplink power consumption is computed based on the uplink airtime and the transmission power [102]:

(6.4)

Algorithm 1 Uplink cellular bit efficiency and airtime computation

Input: ,
,
,
, , , , ,
Initialization:
Compute WRR weights , initialize airtimes and counter :
———,
, [current computed values of airtimes]
, [values computed in the previous round]

Procedure
Iterative procedure (at most loops).
1: **while** and , **do**
First, update bit efficiency:
2: **for** **do**
3: Compute using (6.1)
4: ; [bit efficiency cannot exceed]
5: **end for**
Second, update resource allocation (symbols are assigned to users proportionally to ; unused resources are redistributed iteratively):
6: **for** **do**
7: ; [available uplink symbols/s]
8: —, ; [initialization unsatisfied demand of in symbols/s]
Iterative resource allocation:
9: **while** **do**
10: ; [symbols/s allocated by eNB in this WRR round]
11: **for** **do**
12: ——— ; [symbols/s assigned to]
13:
14:
15: **end for**
16:
17: **end while**
Update airtime values:
18: **for** **do**
19:
20: Compute with (6.3)
21: **end for**
22: **end for**
23:
24: **end while**
Output: ,

where is the power consumed by the terminals when the transmission power is , is a proportionality factor (in), ——, and is the power consumption of the terminals when no transmission is in place. It is worth noticing that the power consumed by a device scales logarithmically with the transmission power, while linearly with the airtime.

6.2.2.2. Downlink

The downlink transmission power of an eNB is constant over time and equal to . The bit efficiency depends on inter-cell interference generated by the transmissions of other eNBs, which are always active. As a result, the average bit efficiency achieved by eNB in the transmissions to user , can be computed as:

(6.5)

where represents the path loss among eNB and user . Note that, differently from the uplink case, the bit efficiency in downlink is not affected by the user airtime, and so it can be computed without iterations.

Downlink resource allocation follows the same VFS scheduling policy used in uplink, so that an expression equivalent to (6.2) holds by replacing the superscript with . So, to compute the downlink user airtime and throughput, we only need to compute the amount of unsatisfied demand , using the WRR policy for resource allocation. For such computation we use an algorithm similar to Algorithm 1 in which the "while-do" statement is not necessary. In fact, a single iteration (equivalent to lines 6-22 of Algorithm 1) suffices to return the downlink airtime of each user, from which we compute the downlink throughput:

(6.6)

Finally, the power consumption in reception by is:

(6.7)

where is the power consumed by the terminal while receiving and is consumed when the terminal is idle.

6.2.3. Analysis of 802.11 Operation

In order to compute the throughput and the power consumption achieved in , we analyze a generic clique. We first show how to compute the bit efficiency of transmitters. Second, with the bit efficiencies we propose a new model to compute the throughput based on the ideas introduced in Chapter 5: operations are going to be analyzed by system *states*. We then compute the duration of the states, which is key to estimate the power consumption.

6.2.3.1. Bit Efficiency

The bit efficiency achieved by each PoA , which in channel is computed in , is univocally determined by the particular destination and is computed as follows:

——————— (6.8)

where is the fixed transmission power of and is the path loss from to . Eq. (6.8) does not include interference because successful transmissions in a clique occur when only one device transmits. Since serves different terminals experiencing different path losses, the average bit efficiency of depends on the distribution of transmissions from to the terminals attached to . Indeed, differently from what assumed in state-of-the-art proposals, the average bit efficiency of a transmitter is neither an input of the problem nor a constant, since it changes with the traffic distribution across terminals served by , and the traffic distribution changes with the number of devices in the clique and with their demand. For this reason we present here a novel approximation method to compute throughput and power consumption.

As in Chapter 3, the basic idea behind our model is that, in each channel, operations can be split in intervals of time during which the set of active (backlogged) devices remains unchanged. In each of such intervals, active APs, GOs, MNs and GMs can be regarded as a saturated system for which the analysis of Bianchi for a clique [10] can be used. Therefore, each interval corresponds to a *state* of the system, for which we can compute bit efficiencies, throughputs and airtimes.

Let us consider the PoAs . In a generic state , the terminals with packets to transmit to are denoted as the set , whereas is the set of terminals with packets to receive from . Now, while the average bit efficiency of terminal is computed with an expression similar to (6.8), the bit efficiency of PoA is the average of its bit efficiencies experienced when transmitting to its attached terminals in RR in state (packets have the same average length, otherwise per-user average packet sizes shall be used as weights in the following expression):

—— (6.9)

6.2.3.2. Throughput and State Duration

Due to the packet-fairness of the MAC protocol assumed, all transmitters achieve the same throughput (in packets per second). Having assumed a common average packet lengths for all transmitters, all throughputs in a clique are statistically identical. Such common value is the throughput computed in state with Bianchi's formula [10] (see Eq. (5.2)), in which the average transmission time is computed with the bit efficiencies derived as described above. Note that accounts for the presence of *all* active PoAs and for their attached terminals active

in state , so it expresses the coupling between different PoAs using the same channel in the same area. Moreover, the throughput of a PoA is fairly shared between its MNs or GMs by the RR scheduling policy adopted. Therefore uplink throughputs and downlink throughputs of the MNs (or GMs) of , in state , are as follows:

(6.10)

We now need to compute the time spent in each state to compute throughput and airtime of devices. Unfortunately, the underlying Coupled Processor System (CPS) model for transmissions we previously introduced, and whose ideas are at the base of the modeling presented here, does not allow us to compute the amount of time spent by the devices in a given state, but only a safe estimation of the average throughput each device can achieve. Therefore, we slightly modify what presented in Chapter 5. First of all, we note that the order in which states are visited is not important, because the arrival processes are independent. So we study the evolution of the system among states reordering the intervals of time where the same set of devices are active. We start with the state where all devices are active and all the queues at the APs/GOs are backlogged. We name such state : the system remains in such state until the uplink or downlink demand of one device is completely served. The portion of time the system stays in state is:

(6.11)

After , the system switches to state in which one less downlink/uplink flow is active, which changes the average system throughput. Thus, the computation of and of the durations of successive states can be done sequentially, until all demands are served or . The values represent the fractions of time spent in each state. The resulting average throughputs on are therefore as follows:

(6.12)

Note that the approach proposed is an approximation, since it assumes that after a flow activates or deactivates, the system reaches instantaneously its steady state, in which Bianchi's formula holds. However, we will show via simulation that this approximation introduces negligible errors.

6.2.3.3. Airtime and Power Consumption

In order to compute the power consumption of APs and MNs (or GOs and GMs) over , we exploit the results of [103]. Specifically, throughput and airtime determine the power con-

sumption of PoAs ( ) and terminals ( ):

—— —— (6.13)

—— —— (6.14)

where is power consumed by an device when idle, and are the powers consumed by a device per airtime unit in transmission/reception, and are the powers consumed per packet transmitted/received per second, and is the average packet length. In the formulas, we have denoted by and the transmission/reception airtime of , whereas and denote airtimes at PoA .

We have shown how to compute the throughput in the previous subsection, we now show how to compute the airtime as an average over states . We denote by and respectively the probability that a system slot (which has variable length, as defined in [10]) contains a transmission and the conditional probability that a transmission is successful in state , given that the transmission occurs. Using the same approach as in [10], we can compute and from the number of transmitting devices (PoAs and terminals) in state , which we denote by :

(6.15)

(6.16)

——————————— (6.17)

where is the the minimum congestion window size of and is the maximum number of transmission attempts for a packet that suffers collisions. The absolute probability to have a success for any of the transmitters is then , so that the probability of collision in a slot for a given transmitter is . Thus, solving (6.16) and (6.17) for , leads to the airtime computation for state :

——————————— (6.18)

———————— (6.19)

————————— (6.20)

—— —— (6.21)

where is the duration of a slot containing a collision,[1] is the successful transmission duration from to PoA , and is the successful transmission duration from PoA to terminal .[2] Considering that unused slots have fixed duration , the average slot duration in state is:

——

(6.22)

Finally, the average airtime values are computed by averaging each of (6.18)–(6.21) according to the state durations computed as described in Section 6.2.3.2. Note that, differently from existing studies, the above equations take into consideration the presence of a PoAs in a clique, and the fact that PoAs serve terminals with different channel qualities.

6.2.4. Network-controlled Coupling

So far we have presented how to compute throughput and power consumption of cellular UEs and GOs and MNs considering fixed traffic demands. However the real demand of GOs depends on what they receive to relay on behalf of their GMs. Indeed, it is possible that the traffic to relay cannot be handled by the GO. In absence of coordination, it is therefore possible that cellular resources are wasted for traffic that can not be relayed by the GOs. A network controller may avoid such problems by acting as coordinator among the cellular network and the groups. The controller needs to estimate the quantity of traffic that GOs can actually relay given the limitation of cellular and resources and enforce a rate limiting on the activity of GOs and GMs. Such control beneficially impacts the resource allocation of UEs that do not act as relays, since more resources are freed for UEs. In addition, since GOs' activity on the channels is reduced with respect to the uncontrolled case, also APs and their MNs indirectly benefit from the presence of a controller regulating the GOs and the cellular devices.

The controller operates exploiting the analysis presented in the previous sections. It first proceeds with the allocation of cellular downlink resources according to all downlink demands, which determines the maximum relay load of GOs. Then the controller finds the corresponding operational point of the 802.11 cliques, using as well the uplink demands of GMs and MNs and the load of the APs. This computation yields the uplink cellular relay traffic demand of GOs, and a corresponding uplink allocation can be computed for GMs. With such allocation, the controller imposes to GMs to send no more than what the GOs can handle to the eNB, and so the 802.11 load changes (freed resources are redistributed to unsatisfied users). Therefore, the operational point

[1] Collisions last as much as an RTS frame if the four-way handshake is used, otherwise they last as much as a packet transmission plus a timeout.

[2] Successful transmission durations include overheads due to MAC ACK transmission and standard inter-frame spaces.

of 802.11 is recomputed, and results in possibly modified 802.11 throughput for GOs. Cellular downlink resource allocation is then recomputed with the new demand from the GOs, finding a final, stable, working point for the network. More in detail:

Step 0: Cellular downlink resources are allocated with the procedure presented in Section 6.2.2.2, the demand of GOs being computed local traffic plus demands of GMs:

UE acting as PoA (6.23)

In the notation adopted in the above expression, we have used the fact that a GO acts as UE for the eNB and as AP for GMs. The controller computes the cellular downlink throughput of any GO (alias PoA). Only part of such throughput belongs to , smaller than , while the rest is for relay.

Step 1: The controller computes the partition of the achieved downlink throughput of a GO, , to the GO itself and its GMs. The partition is computed considering RR and the downlink demands.

Step 2: The controller computes the uplink throughput of each GM using the procedure described in Section 6.2.3, but replacing with the GM downlink throughputs found in **Step 1** as downlink demands.

Step 3: The cellular uplink can then be computed with the procedure of Section 6.2.2.1, using GMs' throughputs computed in **Step 2** as GO relay uplink demand.

Step 4: The resulting cellular uplink throughput of the GO is split across its GMs as in downlink (with RR). Such throughput values are set by the controller as rate limits of the GMs.

Step 5: The controller can now recompute the operational point in for all APs, GOs, MNs, and GMs, with the limits fixed for relay traffic in **Step 1** and **Step 4**. In this step, any GM obtains exactly the rate fixed with rate limiting, while a GO can receive no more than what imposed in **Step 1**.GMs and GOs imposed transmission rates can be relayed to the cellular infrastructure. This step ensures no waste of cellular or resources.

Step 6: GO downlink relay throughputs (in) computed in **Step 5** are used to recompute the downlink resource allocation with the procedure of Section 6.2.2.2.

With the above, we have not only the throughputs, but also the demands of all terminals, accounting for rate limiting enforced to GOs and GMs. So we can also compute airtime and power consumption of all terminals with the methods described in Sections 6.2.2 and 6.2.3.

6.3. Energy Efficient Access Selection

In this section we present a new access selection mechanism that improves the energy efficiency of wireless terminals in mixed access networks. Exploiting the fact that carefully selecting the attachment of the terminals to proper PoAs leads to different throughputs and power con-

sumptions, our access selection mechanism aims to maximize the total terminal energy efficiency figure:

(6.24)

where is the set of UEs (some of which acting as GOs), MNs, and GMs present in the system, and are the throughputs achieved by terminal in uplink and downlink, respectively, whereas is the power consumed by . If is a GO, and do not take into consideration the traffic which is being relayed towards the GMs, while includes the portion of power consumed to receive and relay the traffic of the GMs.

In our access selection mechanism, terminals could access the Internet selecting between eNBs, APs and GOs, under their coverage. Moreover, UEs can become GOs *if needed*, i.e., if there is a benefit in terms of total energy efficiency (6.24). However, purely pursuing the total energy efficiency might lead to low and unfairly distributed throughputs. For instance, an AP might be more energy efficient than an eNB but guarantee less throughput when the number of MNs becomes large and collisions drastically degrade the throughput in . Furthermore, when acting as a GO, a UE may use a non-negligible part of its energy to relay the traffic of the attached GMs. The individual energy efficiency of a GO is negatively impaired by the relay traffic, and being a GO is not beneficial for the UE acting as GO. It can be beneficial for the total energy efficiency though. Keeping GOs in the loop is also beneficial from the point of view of connectivity and from the point of view of the users, since each GO might turn to be a GM at some point in time.

To account for the above considerations, in the following, we formulate an access selection problem that aims to maximize the total energy efficiency under throughput and energy constraints defined on a per-terminal basis.

6.3.1. Problem Formulation

In our formulation, we use the total energy efficiency as utility function, and we add per-terminal constraints to guarantee that the achieved terminal throughput is at least times the highest throughput potentially achieved under any other access choice, , and the energy efficiency achieved by a GO is at least the times the energy efficiency achieved by acting as a simple UE, . Denoting by the energy efficiency of terminal , and by the set of feasible terminal access combinations, the resulting formulation of our optimization

problem is as follows:

s.t.:

(6.25)

In the problem formulation we have split the contribution to the total energy efficiency in terms of MNs (the first summation) and UEs (second summation). The contribution of UEs is further split between non-GOs (UEs) and GOs (UEs).

Due to the fact that, in a real mixed access network, the access selection mechanism has to take separate decisions every time a new terminal enters the network or every time a terminal needs a handover, we actually tackle the above optimization problem in on-line fashion. The resulting on-line problem can be reduced to an on-line Generalized Assignment Problem (GAP) [104], in which items arrive one after the other and they have to be assigned to bins of fixed capacity. In GAP a decision has to be made at the arrival of each item and cannot be changed later. When assigned to a bin, the item requires a given capacity and provides a given benefit. Depending on the bin, capacity and benefit may change. The objective of GAP is to store the *subset* of available items so to maximize the benefit achieved. In our on-line problem each terminal can be seen as an item in GAP, while each of the PoAs is a bin. Depending on the PoA, the capacity required and the energy efficiency (the benefit) achieved by each terminal change. However, our problem is more constrained than GAP, since we need to assign a PoA to *each and every* terminal. Therefore, if the capacity requested by a subset of terminals to a PoA is higher than the actual capacity of the PoA, each terminal has to shrink its request so to fit into the PoA, at the expenses of reduced efficiency. Our problem is a GAP with extra constraints and with bin-occupancy-dependent benefits. Since GAP is NP-Hard, so is our on-line problem.

6.3.2. On-line Energy Efficient Access Selection

In order to find a solution to our optimization problem, we propose the Marginal Benefit Heuristic (MBH), inspired a the well known heuristic that has been proposed for solving

GAP [104].

Upon the arrival of a terminal, or at handover request, MBH works as follows:

1. MBH computes the energy efficiency of each terminal before the attachment of the new terminal.

2. MBH evaluates the attachment of the new terminal to any feasible PoA. For each possibility, MBH computes the energy efficiency of affected terminals.

3. At this point, all the PoAs not ensuring at least of the maximum throughput that could be received by the new terminal are discarded. Likewise, when the PoA is a GO, the MBH algorithm evaluates the energy efficiency degradation experienced by the GO. To have a reference point, all the GMs attached to the GO under analysis are considered as attached directly to their closer eNBs and the GO energy efficiency is computed. If by serving all its GMs and the new terminal the energy efficiency of the GO drops below times the energy efficiency it would achieve while acting as a simple UE, then this GO is discarded from the PoA candidate list.

4. The MBH choses a PoA which guarantees the highest marginal benefit (w.r.t. the same PoA without the new terminal attached) among the remaining options.

Our approach differs from the classic heuristic used for the GAP problems mainly because it does not require to evaluate the allocation of a new incoming terminal to each and every possible in-range PoA. In fact, the additional constraints we introduced, i.e., on throughput and energy guarantees for the incoming terminal, identify a subset of possible target PoAs. Another difference is the fact that MBH, differently from GAP heuristics, accepts negative marginal benefits, since MBH needs to allocate the new terminal to one PoA in any case. Moreover, and most importantly, our heuristic has to recompute the resource allocation for each possible assignment.

6.4. Numerical Evaluation

In this section we first validate our method to analyze performance. The cellular part of the model is quite straightforward so we skip validation results here. Afterwards, we show how our MBH scheme outperforms state of the art approaches and we shed light on the trade-off between maximizing system energy efficiency and achieving fair throughputs.

6.4.1. 802.11 Model Validation

To validate the throughput and power consumption model we presented in Section 6.2.3, we used a packet-level simulator written in MatLab. In our simulations, a GO and a number of GMs randomly picked among 2 and 10 were positioned in a area, following

Table 6.1: parameters

Carrier	
(Min congestion window)	
Max retry	5
Other MAC param. (SIFS, DIFS,...)	
Max Transmission rate	
Noise Level	
Bandwidth	
Path Loss	Log-Distance, Exponent 3
Transmission Power	

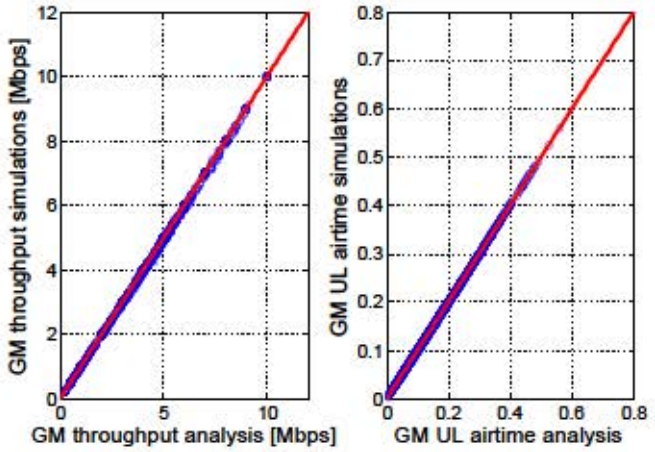


Figure 6.2: model validation.

a uniform distribution. The channel quality among the GO and the different GMs (and viceversa) was computed as in (6.8). The simulator considers that all the GMs and the GO can listen to the transmission of all the others. Transmissions were therefore regulated by the MAC protocol, whose main parameters are show in Table 6.1. Packet arrivals followed Poisson processes with intensity fixed during the simulation and picket uniformly at random in the set

.

In Fig. 6.2 we compare simulated and analytically computed throughput and airtime of terminals. Each point corresponds to a simulated throughput (airtime), whereas the red line indicates where the point should be if the model was error-free. As it is easy to see, the approximation introduced with our state-based analysis is practically negligible, and the largest relative error is for throughput and for airtime. Furthermore, although not explicitly shown in the figure, the percentile of the error on throughput is as low as , while the one for the airtime is .

6.4.2. Access Selection Performance

In this section we evaluate the performance of our access selection mechanism MBH. In particular, we quantify the gains in terms of energy efficiency achieved by MBH against state of the art solutions when the values of and change. We also consider the effects of access selection decisions on aggregate throughput and its fair distribution among terminals.

Table 6.2: Cellular transmission and power parameters

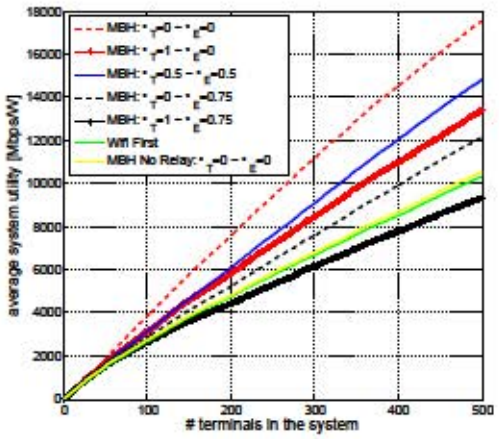
Carrier	
Bandwidth	
Subframe Duration	
Symbols per Resource Block (RB)	
Max Power Uplink	
Power Downlink	
SNR target Power Control	
Max Symbol Efficiency	
Noise Level	
Path Loss Model	Log-Distance, Exponent

In an area of , we first fixed the position of eNBs so to have an inter-site distance equal to m. Furthermore, in each experiment, we positioned randomly APs with uniform distribution. The APs select the particular channel they use depending on their location (the same holds for UEs acting as GOs). In this way, terminals scheduled on the same channel are co-located and form a clique. We further introduce users, one after the other, at random positions. At their arrival, we apply MBH, and we test different configurations for and . The demands of the users are like the ones described in Section 6.4.1, and downlink and uplink demands of a terminal are picked independently. Note that, while adding new users, we evaluate the performance of MBH under different (increasing) density of terminals.

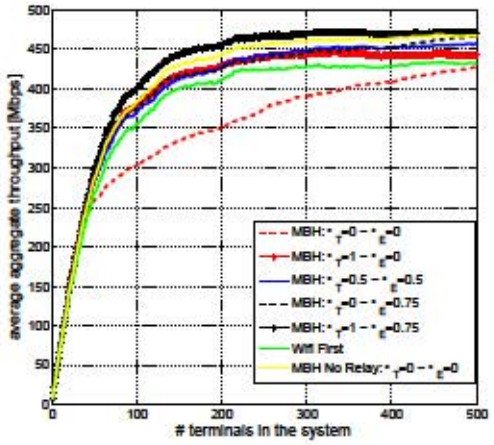
and cellular parameters are set up as shown in Table 6.1 and Table 6.2, respectively. The latter also shows the particular power consumption parameters used, extracted from [94], [103], and [102]. Furthermore, we set to zero. In this way we effectively evaluate only the power consumption of the network cards of the terminals when transmitting/receiving.

Fig. 6.3 presents the results achieved in terms of utility (i.e., energy efficiency), aggregate throughput and fairness varying the values assumed by and . Results are compared with a standard *WiFi First policy* [51], in which a new user is attached to the strongest AP, if available, or to the strongest eNB if no AP is available. Results are also compared with a version of MBH which ignores D2D relay opportunities. The number of independent channels considered in Fig. 6.3 is 3, but similar results have been achieved with more channels.

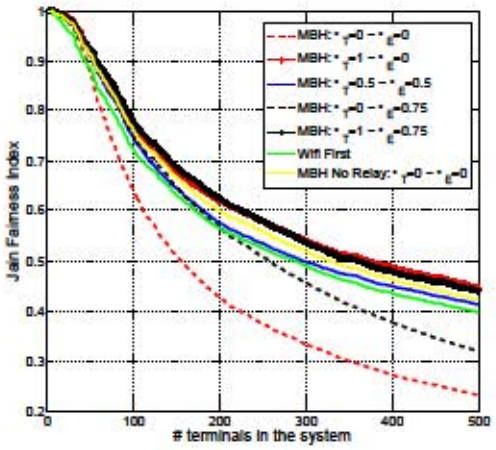
The main tendency shown by Fig. 6.3 is that MBH is able to achieve huge energy efficiency gains when compared to WiFi First policy if there is no guarantee on per-user throughput and GO power efficiency. When and are set to zero, indeed, MBH improves the energy efficiency at the terminal side by about the if compared to the WiFi First policy. However, a careful inspection of the results shows that MBH with and set to zero aggregates the traffic and the energy efficiency to a small number of users present in the scenario. As a result, both the



(a) Utility (energy efficiency).



(b) Aggregate throughput.



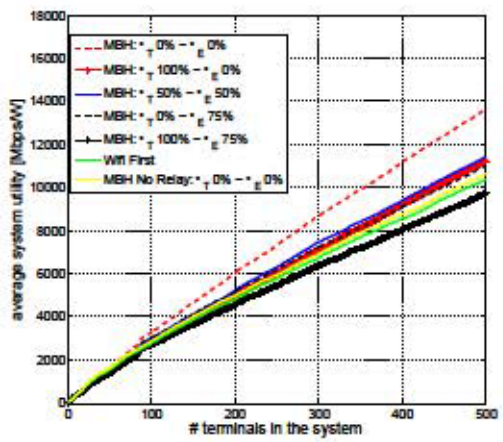
(c) Jain fairness index.

Figure 6.3: Performance of MBH considering different values of α_T and α_E (with 3 channels for 802.11).

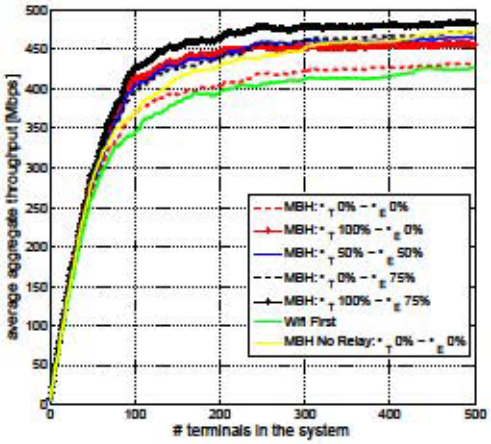
aggregate traffic and the per-user fairness decrease w.r.t. the WiFi First policy.

On the one hand, increasing the value of α_T allows coping with this problem. Indeed, restricting the access selection of the new users to the subset of PoAs ensuring at least α_T times the maximum achievable throughput improves drastically the aggregate throughput and fairness. On the other hand, such result is achieved at the expense of GO energy efficiency, and the figures shows that utility drops when throughput grows. For instance, when α_T is 1 (and $\alpha_E = 0$), a new user always attaches to the PoA ensuring the larger achievable throughput, and MBH energy efficiency gain drops to 35%—which is still remarkable—with 15% higher fairness w.r.t. WiFi First.

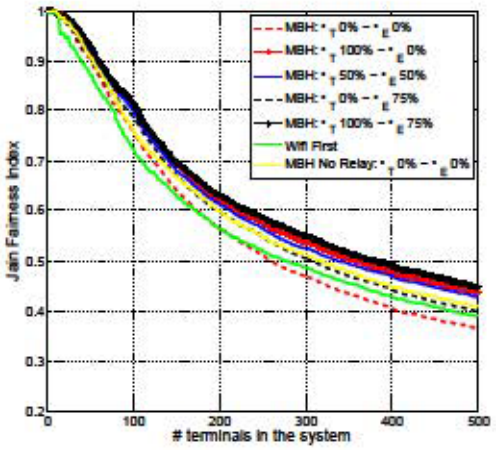
By increasing α_E, we can ensure that GOs limit their D2D relay activity, i.e., we reduce the average number of GOs in the system. In turn, this reduces the achievable energy efficiency gain. Nevertheless, Fig. 6.3 shows that a trade-off between system energy efficiency, fairness and power consumption wasted due to relay is possible. For instance, when we select $\alpha_T = 0.5$



(a) Utility (energy efficiency).



(b) Aggregate throughput.



(c) Jain fairness index.

Figure 6.4: Performance of MBH considering different values of α_T and α_E (with 3 channels for 802.11), when only the 30% of the UEs accept the GO role.

and $\alpha_E = 0.5$, we achieve gain in energy efficiency close to 50%, while achieving competitive aggregate throughput and fairness w.r.t. WiFi First. We also ensure that at most half of the power consumed by a GO is reserved for relay. We also note that D2D is key to enable energy efficient access networks. In fact, when MBH is not allowed to create relay links, performance figures drop and become very close to (though better than) what achievable with WiFi First. Note also that, interestingly, the utility gain is barely affected by the population density, which tells that MBH scales well.

To further evaluate the value of D2D PoAs, we have designed an experiment in which only 30% of the UEs accept to play the GO role. The results, shown in Fig. 6.4 for the same scenarios analyzed in Fig. 6.3, illustrate how the energy efficiency of MBH keeps outperforming the one achieved by legacy access selection schemes (up to the 37% gain in utility). However, with less GOs and less D2D opportunities, the utility (energy efficiency) grows at a slower pace with the network population. The figure confirms that the phenomenon observed in Fig. 6.3, i.e.,

that increasing reduces the utility but also increases throughput and fairness, was due to the reduced number of available GOs. Indeed, in Fig. 6.4, with systematically less GOs, fairness as well as overall throughput increase in all MBH configurations w.r.t. Fig. 6.3. However, this experiment illustrates how even a limited number of D2D opportunities has a high positive impact on energy efficiency.

Part II : Joint Orchestration of D2D Communications and Other Optimization Techniques

Chapter 7

Two-level Opportunistic Spectrum Management for 5G Radio Access Networks

In the first part of this thesis we analyzed the effects of introducing Device-to-Device (D2D) communications within the cellular infrastructure. The analysis aimed at showing the system performance when D2D communications were used as a mean of communication among users or as a relay technique towards the internet. In this chapter, and in general in this part of the thesis, we analyze the effects of D2D communications when they are used jointly with different techniques.

Several techniques have been proposed to independently cope with interference or low spectral efficiency in RANs, such as beamforming, Multiple-Input Multiple-Output Transmissions (MIMO) or many others, as shown in [3]. Leveraging some preliminary promising results published in [66], in this chapter we propose a novel control mechanism that *jointly* deals with interference and spectrum efficiency by coordinating intra-cell and inter-cell resource allocation strategies. We call such a control mechanism Two-level Opportunistic Spectrum Management for 5G Radio Access Networks (TOMRAN).

Specifically, TOMRAN exploits an Inter Cell Interference Coordination (ICIC) scheme based on the Almost-Blank SubFrame (ABSF) paradigm for inter-cell resource scheduling, and an outband D2D relaying strategy for collaborative users for intra-cell resource allocation. ABSF represents a 3GPP standard technique to deal with the interference caused by neighboring cell activity, a.k.a. Inter Cell Interference (ICI). ABSF mechanism prevents the base station (e.g., the evolved Node B (eNB) of LTE/LTE-A networks) from transmitting in a specified set of subframes (as specified in the ABSF pattern information), where only control signaling is permitted. However, even when subframes are blanked, some physical resource elements, such as the Cell-specific Reference Symbolss (CRSs), must be active to provide channel measurements and estimations, hence the term *almost blank* in the ABSF acronym. This causes some interference that can be

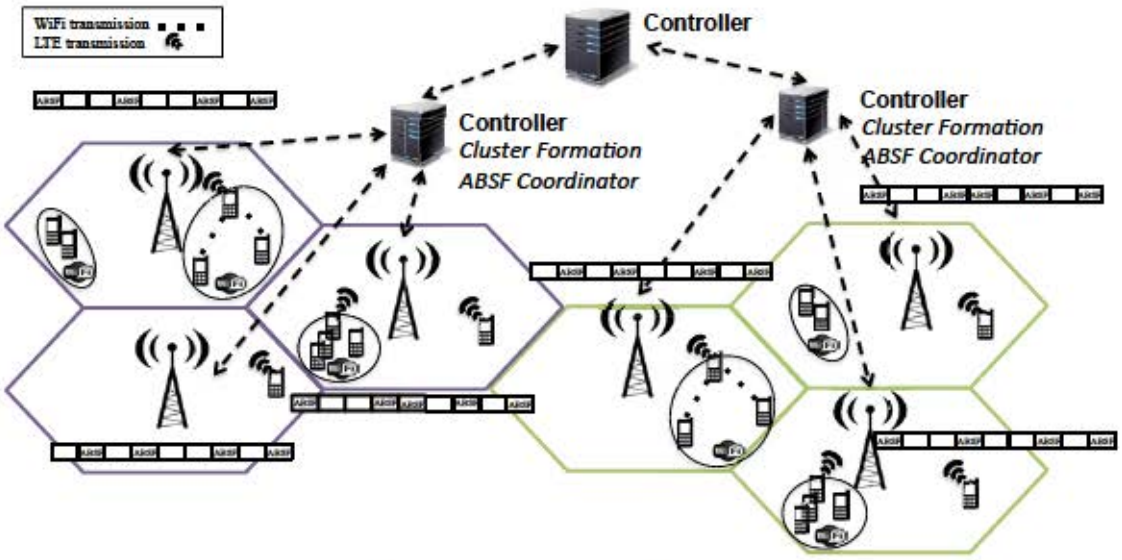


Figure 7.1: Network architecture.

easily handled by interference cancellation techniques, as proposed in [105, 106]. Therefore, ABSF allows to use higher Modulation and Coding Scheme (MCS) and leads to achieving higher throughputs by reducing the overall sensed interference. As we showed in Chapter 4, Chapter 5 and Chapter 6, D2D communications represents instead the key-enabler to cope with low spectral efficiency in the future high-density network design.

Differently from other proposals, TOMRAN takes into account the real capacity attained by each technology involved, e.g., Long Term Evolution-Advanced (LTE-A) [107] and WiFi Direct for relaying [108], and does not neglect their impact on each other's performance. As we already mentioned in Chapter 6, offloading traffic through a contention-based system, such as an -based WLAN, may result in serious congestion that dramatically degrades the overall system performance. Conversely, TOMRAN estimates the stability region of the D2D links and guarantees that the cellular traffic is effectively relayed. Indeed, TOMRAN exploits the results in Chapter 5 in order to correctly evaluate the achievable rates. Making use of such features, TOMRAN dynamically adjusts the amount of the offloaded traffic.

The organization of the chapter is as follows: in Section 7.1 we present TOMRAN. In Section 7.2 we discuss advantages offered and complexity issues tackled in TOMRAN. Finally, we present in Section 7.3 an extensive simulation campaign to quantify the gains provided by a joint inter-cell and intra-cell resource allocation.

7.1. The TOMRAN Control Application

We assume a multi-cell cellular network with frequency-reuse- and users in the network. We also assume that a SDN-based architecture is deployed in the cellular access network, similarly to the one mentioned in Chapter 6. In such architecture, a controller manages a small piece of the network based on fixed policies and adaptive rules. The controller dynamically takes re-

source allocation decisions and promptly issues scheduling policies to the base stations under its control. Fig. 7.1 provides a high level view of this architecture. We envision TOMRAN to be implemented in an application deployed on top of the controller in charge of gathering all necessary information for D2D clustering, D2D intra-cell resource allocation, and inter-cell interference coordination based on the ABSF technique. Operational details are provided in the following, pointing out how a joint intra-cell and inter-cell resource management can be properly achieved. In this context, we analyze only downlink communications, leaving the analysis of uplink out of the scope of this chapter.

D2D clustering and relay. TOMRAN is responsible for the cluster formation process. Each controller in the system collects the information regarding the channel quality and the location of the users. Based on such information, D2D clusters are formed in the network by means of a *Merge & Split* algorithm [21] targeting the maximization of per-user throughput. In particular, TOMRAN takes into account the distance among cluster members and the achievable throughput gain after clustering them, and assumes that the clustering gain is fairly shared among cluster members. Then, TOMRAN elects the cluster member with the highest cellular channel quality to act as the relay between the eNB and other cluster members. Each cluster member can potentially become a relay once its cellular channel quality becomes the highest in its cluster. Nevertheless, the relay selection decision is renewed only when the channel statistics of the cluster members change considerably. Since TOMRAN is aware of clustering decisions whenever a packet is delivered to a cluster member, the eNB simply transmits the packet to its corresponding relay. However, when transmitting to a particular cluster, TOMRAN separates all users into primary, i.e., relay nodes, and secondary, i.e., all other cluster members. At the eNBs, TOMRAN sends packets to the cluster members, via the relay node, only when there are no primary users' packets queued. All in all, TOMRAN maximizes the system throughput and improves the energy efficiency since only the most efficient links are enabled as downlink paths from the eNB to the users, without penalizing those users performing as relay nodes.

D2D resource allocation. TOMRAN disposes of several non-overlapping WiFi channels for intra-cell resource allocation for D2D communications (channels in the GHz bandwidth and several more in the GHz bandwidth, depending on the country). After the cluster formation phase concludes, TOMRAN uniformly assigns the available channels to relay nodes. Hence, relays that are scheduled on the same WiFi channel contend for the same WiFi resources, affecting each other's performance. TOMRAN gathers the statistics of user demands, which are available at the base station side, and the D2D link capacity for each user, which is beforehand reported through cellular Channel State Information (CSI) indicators. Once all needed pieces of information are available, TOMRAN evaluates the set of achievable rates for each WiFi channel by means of the CPS model. In particular, in order to evaluate the set of the achievable rates within a cluster of devices, TOMRAN uses the analysis proposed in Chapter 5.

Whenever the demands of the relay nodes cannot be served (e.g., user demands lie outside the stability region of D2D), TOMRAN optimally selects a set of rates in a per-user proportional-

fair manner, which automatically caps the relay demands. Within the set of achievable rates, TOMRAN selects the ones optimizing per-user proportional fairness through the optimization problem presented in Eq. (5.4). We re-formulate the optimization problem in the following:

$$\begin{aligned} & \text{maximize} \\ & \text{subject to} \end{aligned} \tag{7.1}$$

where and are the maximum throughput that relay node is allowed to retransmit to its relay members and the initial total demand of those relay members, is the set of achievable rates for all relay nodes scheduled in the same WiFi channel, as determined by the Coupled Processor System (CPS) model, is the number of relay nodes transmitting in that channel, and is the number of cluster members under relay node .

Inter-cell ABSF control. After achieving the maximum traffic each relay node could transmit to its relay members, TOMRAN manages the coordination of base stations to enforce an ICIC strategy. Specifically, TOMRAN uses cellular channel information for relays and non-clustered nodes (if any), which are the only active cellular users, and the real downlink demand of each cluster, computed through Problem (7.1). Based on the collected information, TOMRAN instructs each of the eNBs to transmit on a periodic pattern of subframes (ABSF pattern). The ICIC algorithm we adopt in TOMRAN is Base Stations Blanking (BSB), a fast centralized approach [22]. In practice, with BSB, every "decision period" (typically frames), TOMRAN collects user and channel information and issues an ABSF pattern to each eNB through the controller. Such pattern is optimized to guarantee a minimum SINR level to any downlink transmission in the cellular network while using as little as possible the cellular airtime to serve at most the offered traffic demand. We remark that BSB has not been thought to be used in a saturated environment, i.e., with downlink queue which have always some traffic to transmit to users. Furthermore, BSB does not provide any guarantee when eNBs coverage areas differ largely. We generalize the approach presented here in Chapter 8. We refer the reader to [22] for more details on BSB.

In summary, from cluster formation to relay capacity estimation and rate allocation, passing through D2D resource allocation and interference coordination, TOMRAN permits to control and optimize cellular network operation at both inter-cell and intra-cell levels, achieving high energy saving.

7.2. Key Performance Advantages

TOMRAN combines D2D clustering and ICIC techniques, which might pursue partially overlapping or even conflicting objectives. However, as shown in what follows, TOMRAN manages to orchestrate the two aforementioned techniques with scalable complexity, achieving efficient

resource utilization and guaranteeing stability.

First, TOMRAN reduces its complexity when needed. The D2D clustering reduces the number of cellular users scheduled by the eNB (only one user per cluster, e.g., the relay node). The selected relay node is the user experiencing the best channel quality in the cluster. When relay nodes are placed close to the base stations, the useful signal strength perceived is much higher than the interference due to the fact that signal attenuation grows quadratically or more with the distance. In these cases, the additional complexity introduced with BSB for enforcing ICIC might only marginally improve the overall network performance. Thus, both ICIC and relay node selection aim at using channels with limited interference. Nevertheless, the complexity of BSB is dominated by the number of users in the most crowded cell [22], and therefore it drops when the coverage area of each eNB reduces. BSB adds complexity to network operations only when the coverage area is sufficiently large and also relay nodes are far from the eNBs, i.e, only when ICIC is truly needed.

Second, the complexity of BSB may become significant when the number of relay nodes in each cell is small while the number of interfering cells explodes. In this case, ICIC and relay selection work in a quite orthogonal way, leaving to the ICIC scheme the ability of improving consistently the scheduling of relay nodes potentially suffering from higher interference. As a result, the complexity of the ICIC technique used in TOMRAN scales automatically to pursue the optimal achievable gains.

Third, to reduce interference, BSB (as any ICIC scheme) limits the number of transmission opportunities of the base stations. This objective conflicts with the goal of scheduling relays as often as possible, to take advantage of their high channel qualities. In turn, the quantity of traffic that relay nodes can handle is mostly limited by the capacity of the D2D systems. Therefore, rather than using independent optimizations of relay nodes activity and cellular scheduling, TOMRAN jointly solves the two problems and identifies whether the system bottleneck lies in the cellular capacity or in the D2D achievable rates. Specifically, to evaluate whether the traffic received by relay nodes can be retransmitted using D2D, TOMRAN uses a conservative estimation of the rates achievable over WiFi D2D links, as in Chapter 5, and instructs the base station scheduler to never exceed such rates using the proportional fair optimization expressed in (7.1). TOMRAN ensures an efficient and fair utilization of the resources at the base stations, and frees the highest quantity of resources, which allows to serve more users. Obviously, it is possible that some relay node would achieve higher rates than the ones assigned by TOMRAN, which exploits only sufficient conditions for stability. Nevertheless, such conditions always guarantee that the backlogs of relay nodes never explode in a cellular network controlled by TOMRAN.

7.3. Performance Evaluation

Numerical simulations have been carried out to assess the system performance. All simulations are based on a cellular scenario, where mobile users are provided with a standard LTE-A

Table 7.1: System simulation parameters

Parameter	Value
Scenario	Circular area
Cellular downlink bandwidth	20 MHz
Number of Base Stations	7
Number of WiFi Channels	11
Number of users ( )	350
Base Station transmit power	27 dBm
Thermal noise power	-174 dBm/Hz
Relay node selection time ( )	every 2 s
Slow fading, Pathloss model	9 dB, UMa [109]
WiFi relay node transmit power	20 dBm
WiFi bandwidth	20 MHz

interface. Users are entitled to form D2D clusters through a second interface, e.g., WiFi-Direct interface, in order to relay the cellular traffic. We benchmark our TOMRAN mechanism against legacy LTE-A. Moreover, we show how either the ICIC or D2D clustering approach would impact aggregate throughput and spectral efficiency when used in isolation. We label as "BSB" those results in which only ICIC is adopted, and as "D2D" those results in which only D2D clustering is involved.

7.3.1. Simulation Details

Our performance evaluation is based on a MATLAB simulator. We consider a scenario that consists of a seven-cell network with users uniformly distributed over a circular area with a variable radius. Users attach to the eNB with the strongest signal and neither move nor leave the network during the simulation. As a result, D2D clusters are given in input to the simulator, since they do not change after cluster formation, while relay nodes can change based on instantaneous cellular channel qualities, which change due to fading. Out-band D2D communications use 802.11n compliant WiFi-Direct in the GHz bandwidth. When clustering is in place, the controller issues ABSF patterns considering only the relay nodes as input for the eNB scheduler. Each eNB schedules relay nodes using a Round Robin (RR) equal time scheduling policy. As previously mentioned, the traffic of relay nodes is prioritized and it is served before the traffic of cluster members. The total amount of cluster members traffic that is queued is finite and it is shaped by a leaky bucket controller, whose long term rate is given by the solution of the optimization problem in Eq. (7.1). The simulation details, based on the values suggested by the ITU-R guidelines for IMT-Advanced networks [109], are summarized in Table 7.1. In what follows, we evaluate the system performance in terms of average throughput achieved by the base stations, assuming that all users offer the same demand.

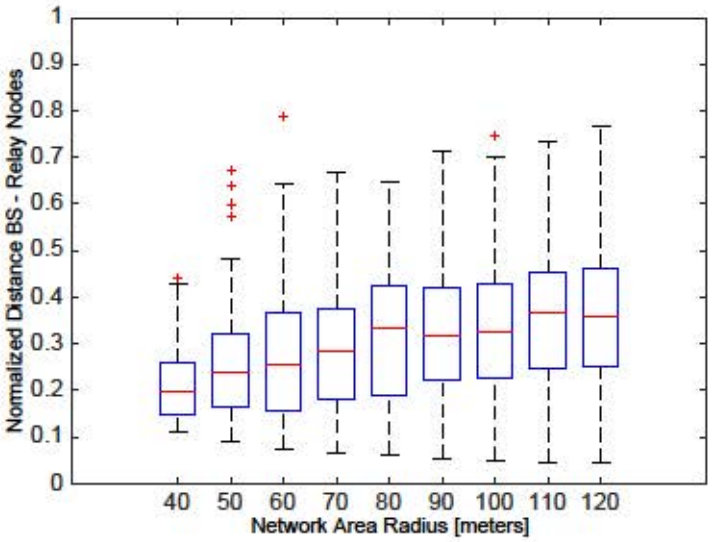
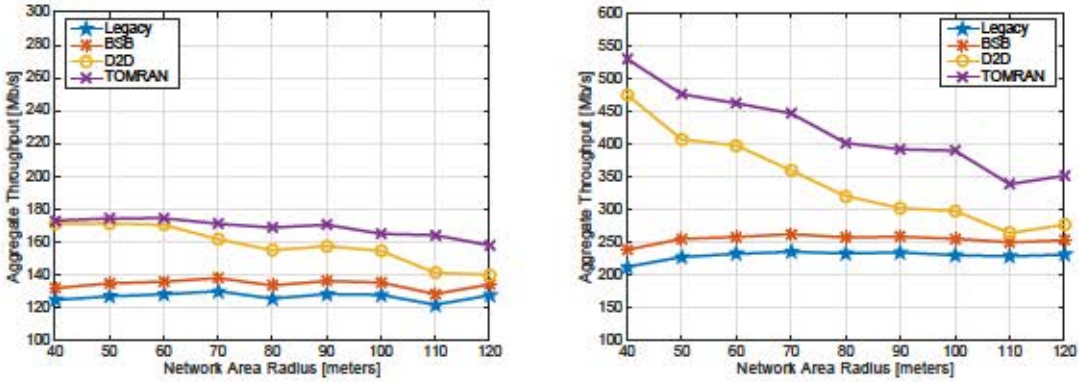


Figure 7.2: Distribution of distances among relay nodes and corresponding BS normalized over inter-site distance (th, th, th, and th percentiles are plotted together with the median, while red crosses represent outliers).

7.3.2. Simulation Results

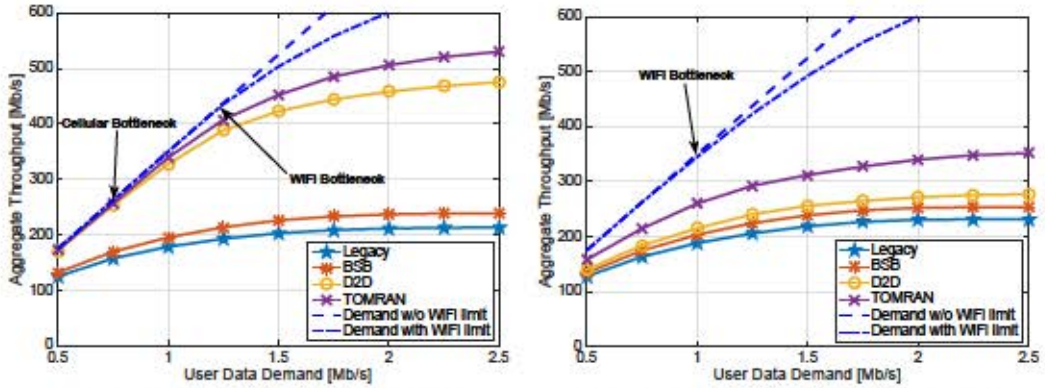
Fig. 7.2 shows the normalized average distance between the relay node and its serving base station as the radius of the simulated area varies. Interestingly, the results exhibit an increasing behavior, which leads to a higher probability to find a relay node on the cell-edge when the network area radius is larger. This is also confirmed by th percentiles reported in the figure, which grow while increasing the network area radius. Thus, the larger the network area radius, the lower the probability to select a relay node with a good cellular channel quality. Even in presence of clustering techniques, this results in a cellular performance degradation, as several cellular users (e.g., relay nodes) suffer from intense interference. Therefore, an efficient ICIC solution is needed to fully boost the system performance by limiting cellular interference effects on relay nodes, as done with TOMRAN.

The network radius plays an important role in the cluster formation process which directly impacts on the performance of the network. In Fig. 7.3(a), we show the throughput of different schemes varying the network radius with a fixed user demand of Mbps, while in Fig. 7.3(b), we increase the user traffic demand to Mbps. Spatial densification of base stations and users (e.g., network area radius equal to meters) seriously impairs the legacy cellular network performance. In such conditions, aggregate throughput always improves even if using only ICIC mechanisms or D2D clustering of users. Specifically, aggregate throughput improves up to when user demands are Mbps and up to when user demands are Mbps. TOMRAN, that jointly exploits the advantages of both techniques, improves aggregate throughput even further. Interestingly, it brings additional gain when the cellular interference becomes challenging (e.g., when relay nodes are on the cell edge). This result confirms our intuitions presented in Section 7.2, showing the ability of TOMRAN to trade off aggregate throughput for computa-



(a) User traffic demand equal to 0.5 Mbps. (b) User traffic demand equal to 2.5 Mbps.

Figure 7.3: Aggregate throughput of 7 base stations evaluated for different network area radius. Simulation parameters are taken from GreenTouch Project.

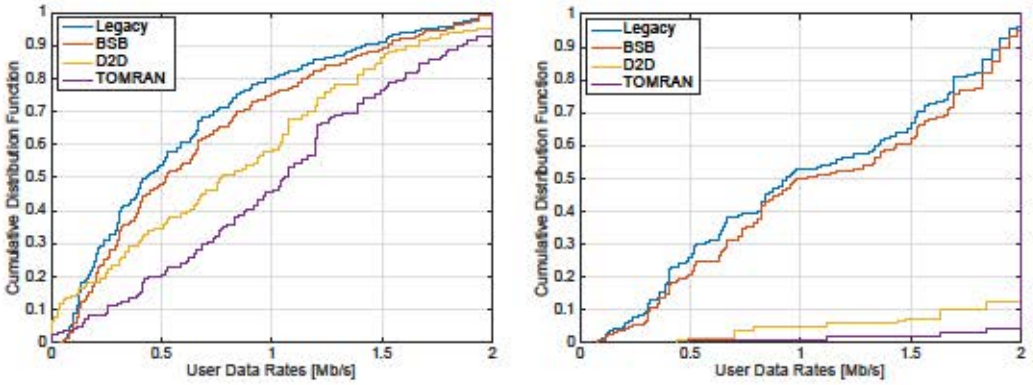


(a) Network area radius equal to 40 meters. (b) Network area radius equal to 120 meters.

Figure 7.4: Aggregate throughput of base stations for different user traffic demands, when users are placed.

tional complexity. Again, if compared with a simple clustering technique, TOMRAN shows an increasing relative gain in aggregate throughput. Indeed, when the per-user demand is Mbps, the gain passes from to when the considered area radius goes from to meters. This validates our intuitions, proving that a compound effect study of applying two independent enhancement approaches leads to a huge gain in term of aggregate throughput, even though WiFi resource sharing limitation is considered.

Fig. 7.4(a) illustrates the achieved throughput for the different schemes tested in a network with radius of meters. The fixed per-user demand ranges from Mbps to Mbps. Notably, we observe a significant improvement with TOMRAN, in which both D2D clustering and ICIC work simultaneously. Indeed, TOMRAN boosts the network capacity by achieving up to gain compared to the legacy scheme. In addition, TOMRAN exhibits up to the and gain in comparison to BSB and D2D schemes, respectively. Fig. 7.4(b) shows the throughput results when the network radius is increased to meters. An interesting difference with the



(a) Cluster members' data rates.

(b) Relay users' data rates.

Figure 7.5: CDF data rates for different schemes in a network with a radius of meters and Mbps as user data demand.

previous case emerges here due to the behavior of D2D clustering when applied either in isolation or jointly with ICIC. The overall gain of TOMRAN and D2D is considerably reduced here, and is comparable to the gain achieved by using the ICIC algorithm in isolation (BSB). This is clearly explained by the user positions and average distances from serving base stations, as shown in Fig. 7.2: several relay nodes are placed at the edge of different cells, experiencing bad SINR values. However, even in this case, TOMRAN outperforms all other schemes by , , and with respect to legacy, BSB, and D2D, respectively. Due to the sparseness of relay nodes in the scenario, D2D-based solutions are indeed not sufficient to cope with spectrum efficiency issues. Thus, simultaneously applying an ICIC mechanism with D2D clustering allows reducing the cellular interference sensed by relay nodes at the edge of the cells and improving the overall throughput. In addition, in both the cases addressed in Fig. 7.4, while the legacy cellular network is not able to satisfy all user traffic demands, TOMRAN is robust to user traffic increasing. Moreover the figures show the demand offered with and without considering the limitation of the WiFi shared media. Note that there are two "special" operational points. The first represents the "WiFi Bottleneck", and shows the minimum per-user demand for which the traffic of at least one cluster member is larger than the D2D out-band capacity. The second represents the "Cellular Bottleneck", and shows the minimum per-user demand at which the TOMRAN capacity does not match the available WiFi capacity. It is important to note that TOMRAN performs similarly to D2D clustering only when all user traffic is served (below the "Cellular Bottleneck"). Moreover, the gain stemming from the coupled control of ICIC and D2D clustering in TOMRAN is not equivalent to the sum of the gains provided by BSB and D2D. In fact, both ICIC and D2D clustering techniques used in isolation rely on wireless channel diversity to improve the performance. However, the diversity gain of D2D after applying ICIC via ABSF patterns is lower because channel quality of users increases due to lower interference (i.e., the opportunistic D2D scheduling gain is lower).

We show in Fig. 7.5(a) the CDF of the average data rates achieved by cluster members, ex-

cluding relay nodes. Even if relay node traffic is prioritized, it is easy to note that cluster members do not starve. On the contrary, they receive a higher throughput than what they would achieve with any other mechanism. Finally, in Fig. 7.5(b) we summarize data rates associated to relay users. D2D substantially improves the relay nodes' condition with respect to legacy cellular networks and with respect to BSB. When TOMRAN is operated, cluster relay nodes' capacity is fully maximized, and high data rates compensate the additional energy required for relaying traffic.

Chapter 8

Stochastic Interference in 5G Networks with eD2D Support

Chapter 7 showed that a cellular access network which implements at the same time Inter Cell Interference Coordination (ICIC) and Device-to-Device (D2D) solutions generates large gains in aggregate throughput, alleviating some of the issues that are present in very dense deployments. However, in this chapter we show that it is very challenging to include ICIC/Almost-Blank Sub-Frame (ABSF) into technically entangled 3GPP specifications as they offer no fairness guarantees in per-user throughput. Fig. 7.5, for example, shows that TOMRAN delivers very different per-user throughputs to cluster members. Furthermore, Base Stations Blanking (BSB) does not provide any guarantee when evolved Node Bs (eNBs) coverage area differs largely.

This is the consequence of imposing (*almost*) interference-free transmissions at the expenses of drastically limiting the number of transmissions in neighboring cells. Furthermore, ICIC/ABSF are PHY-layer techniques wherein throughput fairness is naturally out of design scope. However, we show that ABSF can be seen as a novel tool for scheduling base station activities, hence can be used to enforce fairness among cells, though at the expenses of system throughput. Indeed, we found the use of ABSF to achieve fairness quite intuitive, even though it was not designed for fairness purposes.

In order to cope with the loss of throughput incurred using ABSF only for fairness purpose, in this chapter we assume that short-range *opportunistic cooperative relay* techniques are in place, like in TOMRAN. In the following, we adopt a technological solution ready to serve the next generation of D2D relay communications: mm-waves communications facilities targeted by the IEEE ad standard, commonly known as Wireless Gigabit (WiGig) [23]. WiGig achieves virtually unlimited speeds with respect to the cellular capacity, thus making enhanced enhanced D2D (eD2D) relay groups a solution one step beyond classical D2D. In the following, we use the term eD2D to refer to our mm-waves D2D proposal. The compound impact of ABSF and eD2D relay consists in regulating the number of simultaneous cellular transmissions while at the same time reducing the presence of vulnerable users experiencing poor channel conditions.

Hence, fairness and transmission efficiency can be jointly targeted by simultaneously controlling inter-cell activity (with ABSF) and intra-cell packet relay (with eD2D) within a novel unified framework. To fully understand the potentials of ABSF and eD2D in combination, we derive a theoretical analysis. Specifically, we present a theoretical study on the limitations of ABSF and on the potentials of eD2D, derive stochastic conditions to show how ABSF can be used to steer user fairness, formulate novel and convex optimization problems to set stochastic ABSF activity patterns by leveraging the advantages of eD2D relay and show that the joint operation of stochastic ABSF and eD2D is practical and brings dramatic gains with respect to state-of-the-art solutions.

The rest of the chapter is structured as follows. Section 8.1 presents the framework we focus on. In Section 8.2 we derive a novel analytical model to study the network behavior. In Section 8.3 we exemplify the impact of ABSF and eD2D on a realistic network topology. In Section 8.4 we formulate two new problems for the stochastic optimization of ABSF in presence of eD2D relay, respectively under static and dynamic user density conditions. In Section 8.5 we validate the model and report on performance evaluation.

8.1. eD2D-assisted ICIC Framework

We consider downlink transmissions in a cellular access network with a set of interfering base stations, operated on the same frequency band by the same operator. Users are provided with multi-RAT connectivity, i.e., LTE-A and ad physical interfaces. Base stations implement ABSF, and users can decide to group and let the base station elect the relay user for the entire group. The set of groups will be denoted by , and the size of group will be denoted as . Intra-group relay transmissions use WiGig, and the relay node is opportunistically changed according to whom is experiencing the best channel condition, similarly to what implemented for WiFi-Direct in [56]. We use mm-waves eD2D to refer to intra-group D2D relay. We assume that all groups always have packets to receive, i.e., the downlink queue of each group is saturated.

In such mm-waves eD2D-assisted cellular framework, depicted in Fig. 8.1, we propose a solution that retains the key strengths provided by eD2D and ABSF. However, differently from standard applications, we design a practical scheme to tune the use of ABSF *stochastically*, to ensure user fairness rather than to reduce inter-cell interference, while at the same time counting on eD2D to boost the system throughput by means of packet relay.

8.1.1. User Groups and Mobility

The formation and presence of groups of users leaning toward cooperation is key for the success of 5G opportunistic relay approaches. We target groups of users socially interconnected, the so-called "persistent groups", which may lead to the formation of relay opportunities lasting a sufficiently large period of time. Indeed, we are interested in creating groups among users showing a strong correlation in their pattern mobility, maybe also originated by the fact that

they are sharing the same means of transportation. Thus we use the Reference Point Group Mobility (RPGM) model [110] for our analysis. We define a center of gravity for the group, which imposes group motion parameters, such as speed or direction. Groups are therefore intrinsically present in our access network and how to form groups is out of the scope of this contribution.

To pursue transmission efficiency and system throughput, we allow users to form groups and use a group member as relay node to talk to the base station. For the sake of fairness, groups are served by the base station using a weighted round robin policy assuming the number of users within the group as group's weight. Therefore, the average number of resources allotted to each user remains constant considering all possible group configurations.

8.1.2. eD2D Relay Technology

As for the D2D technology, we propose IEEE ad, commonly known as WiGig. In addition to conventional b/n, where and GHz frequencies are involved, WiGig adds a new frequency band over GHz for ultra-high data speed. In WiGig, a control/user plane network functional split is applied: whenever the high-performance connection is available, a mm-waves connection is established. We propose WiGig for implementing D2D relay, resulting in what we denote as eD2D relay, for two main reasons: as other -based solutions already proposed for D2D, WiGig does not interfere with cellular transmissions, and WiGig achieves unprecedented speeds[1] with very limited interference issues. In particular, up to 32 antenna elements can be grouped together in a very limited space, to perform beamforming for highly directive communications, thus reducing the interference caused by concurrent transmissions. However, the directivity in the communication does not bring any additional delay during association and discovery procedures, as the control plane is managed through conventional channels.

8.1.3. Opportunistic Data Forwarding

With the ultra-high capacity of eD2D, the system throughput bottleneck is represented by the cellular transmission rate only. Therefore, the more efficient cellular transmissions are, the higher the system throughput grows. Thereby it is key not only to select relay nodes opportunistically, but also to be able to switch relay nodes swiftly, as soon as channel conditions vary. We assume that the relay node is selected on a per-packet basis so that cellular transmissions always occur on the strongest cellular channel, like suggested and experimentally validated in [56]. Such continuous re-election of relay nodes has no practical drawbacks from the point of view of end-to-end traffic flows management. In fact, relay decisions are taken on a per-packet basis by the base station (who knows which group member has the better channel), after which, relay nodes serve the aggregate group data demand via WiGig, thus experiencing no practical delay. Therefore, in our system, data forwarding is transparent to protocols such as TCP or UDP.

[1] The n capacity is upper-limited to Mb/s under the best channel conditions, while WiGig currently supports up to Gb/s, but it can ideally reach G/s with x MIMO and QAM modulation.

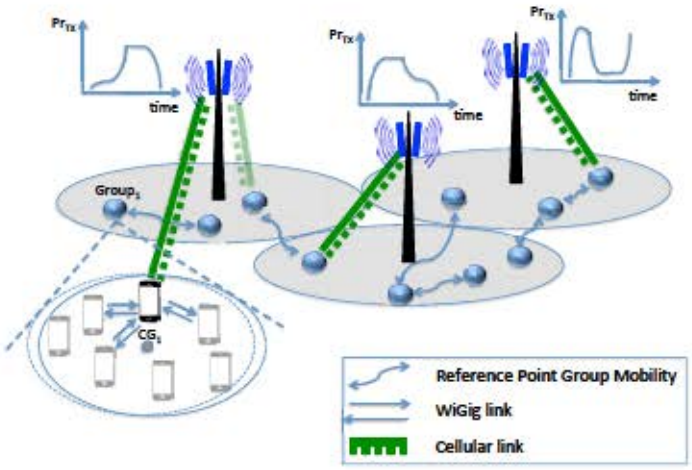


Figure 8.1: eD2D-assisted cellular framework.

8.1.4. ABSF for Fairness

ABSF is a 3GPP time-domain scheduling scheme to prevent macro base stations from transmitting data in a particular set of subframes. Blanking decisions are based on interference suffered in the system, and result in a binary pattern that specifies whether a subframe has to be blanked or not (*ABSF pattern*). Based on the 3GPP standard, the ABSF scheme can be implemented in a conventional cellular system without imposing any constraint on the specific set of subframes to blank. Specifically, standard guidelines propose an ABSF *application ratio* as a number of used subframes over the total number of subframes within a pattern. Once this fixed ratio is imposed by a *network controller*, base stations may make random choices to select the specific pattern of subframes to be blanked.

In order to assess the impact of ABSF, we will compute the transmission efficiency for each possible combination of active base stations. In such computation, the operation of eD2D-relay groups is taken in consideration. We name each of the possible combinations of active base stations as *ABSF state*, and denote it by . We use for the set of base stations not blanked by ABSF in state , whereas we denote the set of all possible states by , with .

By muting subframes for a subset base stations, ABSF helps to transmit at higher Modulation and Coding Scheme (MCS), but it cannot help users with inherently poor channels. In fact, ABSF alone can even lessen the overall volume of traffic served in the cellular network if not properly tuned, simply because some cells are prevented from transmitting.

We go beyond existing schemes, and apply the ABSF paradigm to eD2D-enabled networks in which bad channels are used as little as possible and the number of users considered in blanking decisions reduces to the number of groups. Therefore, with our proposal, we soften the burden of making ABSF decisions. How often a relay node will receive data is a consequence of which ABSF pattern has been selected by a *network controller*, which stochastically builds the patterns to be distributed to the base stations. This can be adjusted over time to pursue not only system

throughput but also fairness.

As analytically shown in Sections 8.2 to 8.4, a *stochastic* throughput-efficient and fairness-optimal strategy for ABSF can be derived by leveraging the impact of opportunistic relay of cellular data traffic on the transmission efficiency achieved under different ABSF blanking patterns.

8.2. Model Design

We analytically evaluate the transmission efficiency and the system throughput performance achieved when cellular data forwarding leverages our opportunistic cooperative eD2D mechanism. Specifically, we compute the transmission efficiency of an *eD2D relay group*, i.e., the average number of bits per symbol that could be transmitted to a cooperative relay group using all cellular resources. We further model the system throughput, i.e., the overall volume of traffic served in the cellular network, assuming that the position of all relay groups in the scenario is known. We finally show system throughput variations upon movements of the relay groups occur. The overall model provides a useful guideline on the system performance optimization.

8.2.1. Transmission Efficiency

Interfering cellular transmissions may cause a severe performance degradation. User channel quality is strongly affected by transmission activities of surrounding base stations.

Let us denote by the average SINR for user in group at location . Given the vector of locations for the users of group , the average SINR can be derived based on the set of active base stations and the propagation model described in [111]. We compute the *transmission efficiency* for each group as the average number of bits per physical transmission symbol downloaded by each group relay node. We can argue that the transmission efficiency of a group determines the resource utilization once group member locations are given. In fact, represents the group capacity normalized to one transmission symbol, given the coordinates of all group members.

Computing transmission efficiencies depends on the position and number of users in the group, as well as on the mapping between SINR and MCS (for further details on MCS mapping examples we refer the reader to [58]). Depending on the distribution of the SINR, we achieve much different transmission efficiencies even with the same average SINR. Furthermore, we should take into account the fact that the relay node changes depending on who is the terminal experiencing the best channel quality. All in all, the transmission efficiency of group can be computed as follows:

$$\text{SINR} \quad \Pr \quad \text{SINR} \quad \text{d} \qquad (8.1)$$

where are the bits transmitted per symbol using MCS and $_{\text{SINR}}$ is the distribution of the SINR experienced by user , whose average is . In the integral we account for the number of bits per symbol transmitted by the users on a discrete set of MCSs as suggested by the standard [58]. Therefore, we cast the function $_{\text{SINR}}$ in a continuous subset of values comprised between and , representing lower and upper SINR levels, respectively, for assigning the MCS to the user . We express the probability for being the relay node of group , as the probability that any other user of group has a SINR lower than the one experienced by user .

As pointed out in [112], SNR experienced by a user at any given location in a urban environment follows a negative exponential distribution (whereas the instantaneous signal follows a Rayleigh distribution) whose average value only depends on the pathloss effect. Radio attenuation and fading are both taken into account in our analysis. The interfering effect of neighbor signal propagation operating on the same frequency is equivalent to a constant noise, since all neighbor base stations are synchronized on the subframe boundaries and use constant downlink power. Hence, the SINR can be also approximated according to an exponential distribution with expected value equal to the ratio between the average useful power received and the sum of the average power levels received from interfering base stations. In the following, we keep this assumption. Nevertheless, the distribution of the SINR can be adapted depending on the scenario and, in a real set-up, can be also experimentally computed. When the SINR respects an exponential distribution, then the following proposition holds.

Proposition 1. *Given the average SINR of user in group located at positions , the transmission efficiency for group is computed as:*

$$\frac{\text{—} \quad \text{—} \qquad \text{—} \quad \text{—}}{\text{—} \qquad \text{—}} \tag{8.2}$$

where is the number of bits per symbol transmitted when MCS is used, is the set of available MCSs, and represent the upper and lower SINR level used to select MCS , represents the power set of the users in group when user is taken out, and is a generic element of such power set (so that is, in turn, a set of users).

Proof: Applying the exponential distribution to Eq. (8.1) is:

$$\text{——} \quad \text{—} \quad \text{d} \tag{8.3}$$

Expanding the product in the integral, we can obtain the following result:

— —
——————————d (8.4)

where represents the power set of , i.e., all the possible subsets of group members that can be obtained from the set when user is taken out, while is one specific subset in . Solving the integral gives Eq. (8.2).

Using the group mobility model described in Section 8.1.1, we can conceive each group as a single point wherein user positions collapsed in the group's center of gravity. Given the coordinates of the center of gravity of group , we can compute the average SINR in this point and assume . Under this assumption, we can readily derive an approximated and simplified expression for the transmission efficiency of group from Proposition 1, as follows:

— —
———————————— (8.5)

Please note that in either (8.2) or (8.5) benefits from applying opportunistic eD2D relay, i.e., the efficiency is computed on cellular transmissions towards the *instantaneous* group leader, which is the node with the highest *instantaneous* SINR value in the group. Therefore, the expression for describes the average of a random variable that is function of negative exponential random variables that are independent but not necessarily identically distributed. The base station will opportunistically select and instantaneously transmit to the group member experiencing the best channel quality, even though all group members might experience the same (or very similar) average SINR (e.g.,).

With ABSF, each state determines interference, and hence transmission efficiency changes from state to state. We denote by the transmission efficiency of group in state , which can be computed with either (8.2) or (8.5) by considering only the set of active base stations. Of course, when group 's base station is prevented from transmitting in state .

8.2.2. Instantaneous System Throughput

We analyze the overall network performance in presence of multiple groups sharing available resources. As mentioned in Section 8.1.1, each active base station implements a scheduler such that each group receives resources proportionally to the group size, i.e., for any group under the coverage of an active base station at time , the resources allocated can be expressed as follows:

$$\text{sym} \frac{}{\text{covers groups and at time}} \quad (8.6)$$

where $_{\text{sym}}$ is the total available number of symbols per second at the base station serving group at time . Please note that is independent of the particular state of the ABSF if

the system is in saturation.

With the above, the resulting instantaneous throughput (in bits per second) achieved by group in state at time is computed with the following expression:

(8.7)

The corresponding instantaneous system throughput in state at time is .

8.2.3. Asymptotic Performance

Let us now consider the impact of group mobility on asymptotic performance, since it directly affects experienced SINR levels. The objective is to compute the mean throughput of group , averaged over time and space, when a given state of the ABSF is considered. We rely on the knowledge of the asymptotic distribution of group locations and we indicate the spatial distribution of the gravity center of group over the scenario as , considering that the movements of the overall group can be approximated by the mobility of its gravity center. Such distribution can be easily derived for well-known mobility models, e.g., the Random Way Point (RWP) mobility model [113] that applies to the gravity center's position if groups follow RPGM, and it can also be empirically retrieved if needed.

Based on the spatial distribution of the group gravity center, and fixing the ABSF state to a particular state , the *conditional* average transmission efficiency achieved by a group under a particular base station , namely , can be computed as follows:

(8.8)

where is the area covered by base station .

Given , the average throughput (over time and space) achieved by group in state can be computed as

(8.9)

where represents the average number of symbols allocated to group under base station , and is the probability for group to be scheduled by base station and it is expressed as

The average number of symbols per second to be used in (8.9), i.e., , is obtained by considering all possible combinations of groups that fall under the coverage of base stations , as follows:

(8.10)

where is the power set of the groups in when group is taken out (is therefore a set too). The calculation of assumes a proportional resource scheduling based upon group sizes, i.e., proportional to the number of users building up the eD2D groups. Interestingly, can be also computed as the time average of resources received from group while scheduled by base station , i.e.:

(8.11)

where only if the position of group at time is served by base station .

The distribution of resources expressed in (8.10)—and therefore the asymptotic system throughput in state expressed in (8.9)—is strictly dependent on the set of active groups and their movements. In case of homogeneous scenarios wherein all groups experience the same spatial distribution of their gravity center, i.e., , so that , , and all groups show the same number of users, i.e., , , is the same for all groups and can be simplified as follows:

(8.12)

The asymptotic system throughput achieved under ABSF state is simply given by the sum of group's asymptotic throughputs, i.e., . The same result would be achieved for the asymptotic system throughput by directly integrating over time the sum of values given by (8.7). However, as ABSF states change over time, the actual total throughput achieved by each group—and hence the fairness level experienced in the network—depends on the fraction of time spent in each ABSF state. We will study the optimization of such fractions of time, i.e., the optimization of ABSF patterns, in Section 8.4. Before that, we proceed by providing the reader with a concrete example that shows how eD2D and the choice of ABSF state impact on throughput and fairness performance.

8.3. Example of Impact of ABSF and eD2D Relay

To evaluate the impact of ABSF on an eD2D-enabled network, we take a realistic heterogeneous dense-urban area of m m close to the Oxford Circus metro station in London city (United Kingdom). Based on its real topography, in the example presented in what follows, we consider only the base stations under the control of O2 mobile network operator.[2]

[2] All information are retrieved from the OFCOM reports available at `http://stakeholders.ofcom.org.uk/sitefinder/sitefinder-dataset/`

Figure 8.2: O2 deployment in London city - Oxford Circus.

In the considered area, base stations are present, as illustrated in Fig. 8.2, which also reports the heterogeneous transmission powers of the base stations. Accordingly, we plot in Fig. 8.3 a snapshot of the group throughputs obtained in three different ABSF states for a static allocation of groups, whose centers of gravity are reported as white dots in the figure. Fig. 8.3 shows the average of the group throughputs over the entire network (top of each subfigure) and over the area of each cell (indicated next to the cell center) when group sizes are , i.e., without eD2D relay, and for all groups, respectively.

As shown in Figs. 8.3(a) and 8.3(b), with all base stations active (i.e., as in the case no ABSF was enforced) the distribution of group throughputs is unfair and low in a realistic deployment, with especially large cells guaranteeing poor throughputs. The effect of eD2D relay is notable and results in a increase in the average of group throughputs over the entire network, though not all cells experience the same degree of benefit.

Figs. 8.3(c) and 8.3(d) show throughputs achieved by keeping active the three strongest base stations only. In this ABSF state, most of the users are under coverage of active base stations, but the interference level remains high. Only the central base station experiences a significant gain (it doubles the throughput for the groups it serves, with and without eD2D relay), since many interfering small cells around it are blanked. Here, the impact of eD2D is less important but still high (on the group throughput averaged over the entire network). Moreover, group throughputs in this state are not only unfairly distributed, but also much lower than in the case with all base stations active. In our experiments, we have observed a similar behavior in many cases (not reported here for lack of space) which questions the ability of ABSF to improve throughputs

in a truly heterogeneous scenario.

Finally, Figs. 8.3(e) and 8.3(f) illustrate how blanking the central (and strongest) cell yields much higher throughputs than in the other considered cases. Therefore, this state is convenient to boost throughput. However, once again, one should notice that many users are not under the coverage of the central base station, and therefore blanked. If used repeatedly, this state would create large unfairness in user throughput. We still notice that eD2D with groups of users brings a significant gain on the group throughput averaged over the entire network.

In general, the distribution of throughputs over groups is very much unfair in any ABSF state and if ABSF states are not chosen carefully, we may incur loss of throughput or user throughputs which are very different with each other. The example of Fig. 8.3 also shows that opportunistic eD2D relay not only boost throughputs, but also attenuates the difference between throughputs achieved under various ABSF states, which means that, with eD2D relay, ABSF can be more likely used to pursue other goals rather than simply transmission efficiency. For instance, since ABSF can be seen as a mechanism that schedules base station activity, it is natural to think of ABSF as of a tool for enforcing fairness by alternating system states conveniently.

The above considerations motivate the problem formulation that will be formally presented in Section 8.4 in terms of optimal probabilities to select ABSF states given the user distribution and given the fact that users help each other with eD2D relay.

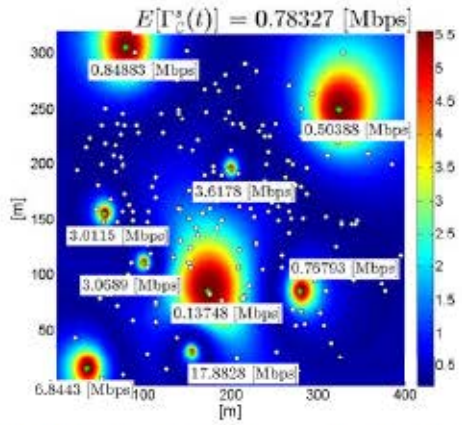
8.4. Proportional Fairness Optimization

As we have showed in Section 8.2, in a real network, both the asymptotic throughput and the instantaneous throughput achieved by the groups change due to two factors: the position of the users, and the ABSF state wherein the system lays.

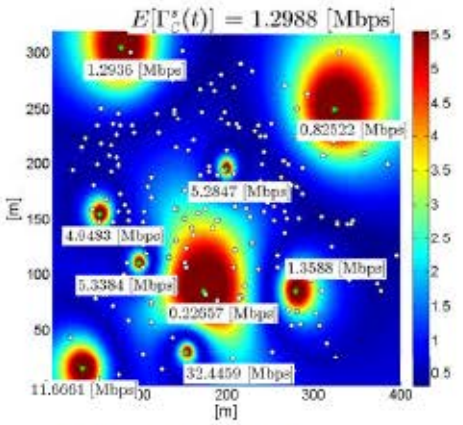
Here we first compute the stochastic ABSF pattern that achieves asymptotic maximal user fairness according to the classical concept of *proportional fairness*, which holds under a vaste set of heterogeneous conditions. Under the assumption of ultra-dense scenarios, such an asymptotic analysis also approximates well the normal behavior of the network, since density conditions do not change over time.

Afterwards, for highly dynamic scenarios wherein user density can fluctuate over time and the ultra-dense assumption cannot be used, we design an easy-to-deploy stochastic ABSF mechanism that jointly achieves high transmission efficiency as well as maximal user fairness, closely following the variations of channel qualities and group locations in the system.

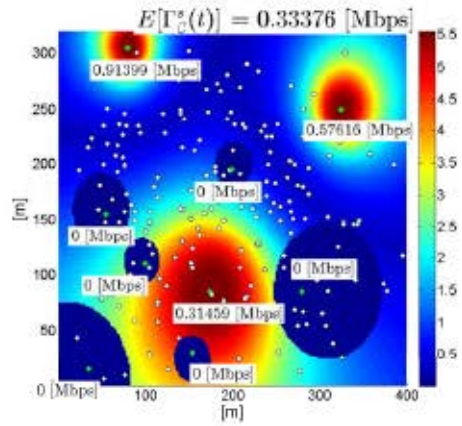
Before proceeding, note that, both with asymptotic and dynamic optimization, implementing the resulting stochastic ABSF patterns has a twofold advantage: random patterns do not incur systematic discretization issues that might arise with deterministic allocations of states, and they make it possible to generate distinct patterns for distinct base stations independently, thus reducing the complexity of *network controllers* issuing the patterns.



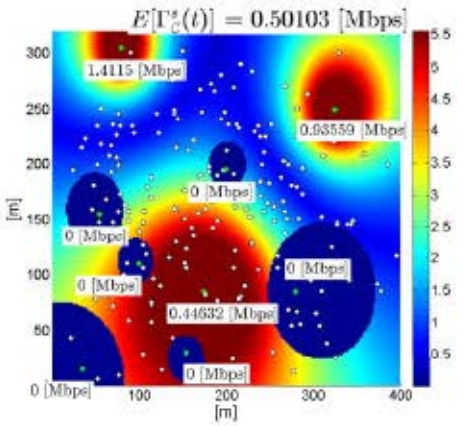
(a) All base stations active, without eD2D ( ).



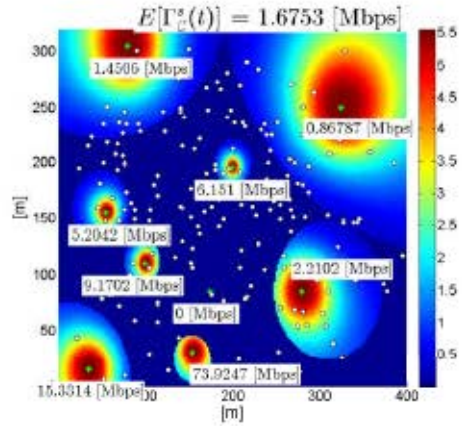
(b) All base stations active, .



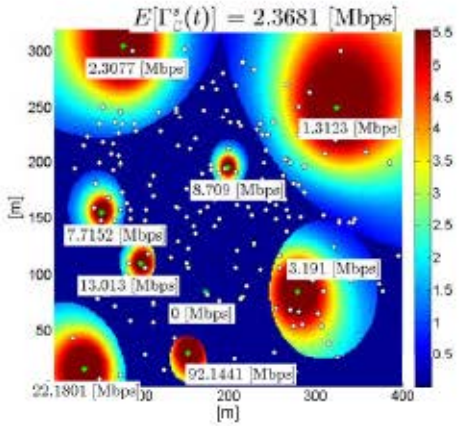
(c) Strongest base stations only, without eD2D ( ).



(d) Strongest base stations only, .



(e) Blanking the central base station, without eD2D ( ).



(f) Blanking the central base station, .

Figure 8.3: Example of throughputs achievable in a realistic network deployment. Figure best viewed in colors.

8.4.1. Asymptotic Proportional Fairness Optimization

Given the ABSF state , we have shown in Section 8.2.3 how to compute the asymptotic throughput achieved by each group in the system, even in heterogeneous conditions. is indeed the number of bits per second transmitted to group , taking into account the particular spatial distribution of the group and all the boundary conditions. The ABSF mechanism allows switching among different states, so to achieve as performance the average of what achieved over the selected states. Specifically, denoting by the fraction of time during which state is enforced, the resulting asymptotic throughput of group is simply given by

(8.13)

can be re-interpreted as the probability mass distribution of ABSF states, and *the analysis unveils that performance is strongly affected by the set of ABSF states used and their probabilities rather than by the order in which such states are visited.* Based on such insight, we stochastically approach the ABSF selection problem, i.e., we assign probabilities to select ABSF states, and use such probabilities to generate sequences (*patterns*) of ABSF states randomly. Hence, proportional fairness between user throughputs is achieved by optimizing ABSF state probabilities.

The optimization problem that we formulate in order to select such ABSF state probabilities goes as follows:

Problem `Asymptotic ABSF-PF`

Select , so to:

maximize

subject to: (8.14)

where weights are used to tune the group fairness. Since the argument of the function in the maximization is linear with the decision variables , the problem is convex and admits a global optimum that can be found with any off-the-shelf solver. Moreover, by linearizing the problem (e.g., the can be approximated by a polygonal chain), the optimum can be found in polynomial time.

Once the probabilities are computed, the node running the optimization (i.e., a *network controller*) stochastically builds and distribute a stochastically optimal ABSF pattern by choosing, for each subframe composing the ABSF pattern, a state at random according to optimal probabilities. The ABSF pattern is then repeated indefinitely.

8.4.2. Dynamic Proportional Fairness Optimization

We next focus on a highly dynamic evolving system and we provide a mechanism to optimize ABSF patterns to achieve proportional fairness over time, accounting for network dynamics as they are observed.

To formulate our proportional fairness optimization problem, let us consider that, in short intervals of time of duration , in which mobility effects are negligible, the throughput achievable in each state by each group does not change and can be indicated as the group throughput computed at any point in time within that interval. Thus, considering a time-slotted optimization framework starting at time , composed by intervals , we can denote the throughput as computed with any value of chosen in . Also in this dynamic version of the ABSF optimization, during intervals , it is however possible to chose subsequently different ABSF states, so to achieve as performance the average of what achieved over the selected states. We denote as the fraction of time during which state is enforced in interval , the resulting throughput of group in interval is:

(8.15)

Clearly, the order in which ABSF states are visited is not important and the computation of such ABSF state probabilities must be repeated every interval , due to network dynamics. The choice for the duration of such interval is pivotal for system performance: a *short* interval allows to consider the network as static, while a *long* interval accounts for including several ABSF states, which in turn increases the accuracy resulting from implementing optimal probabilities with a finite-length ABSF pattern. For instance, the probabilities could be chosen once per second, in line with normal ABSF decision-making procedures, which involves patterns of tens or hundreds of states wherein each state lasts at least 1 ms.

The optimization problem that we formulate in order to select such ABSF state probabilities is based on a long-term proportional fairness metric, in which the throughput is observed over a period of past intervals and predicted for the next interval . To achieve so, the optimization is repeated at the begin of each interval , and we define a utility function based on the of group throughputs (to introduce proportional fairness) computed over intervals:

Problem `Dynamic ABSF-PF`

At time , select , so to:

maximize

subject to:

(8.16)

where weights are used to tune the group fairness target and coefficients define how past

samples of throughput affect future decisions. Since , which is the only unknown term in the sum inside the argument, is linear in the decision variables , also the above-defined dynamic version of the optimization problem is convex, admits a global maximum and can be easily linearized and solved in polynomial time. Every time probabilities are computed, the network controller stochastically builds and distribute a new ABSF pattern by choosing, for each subframe composing the ABSF pattern, a state at random according to new optimal probabilities. Such ABSF pattern is repeatedly used until a new pattern is issued.

8.4.3. Remarks on the Stochastic Optimization of ABSF Patterns

In both Problems `Asymptotic ABSF-PF` and `Dynamic ABSF-PF`, weights can be selected based on the desired fairness target. E.g., for targeting equal throughput on a *per-user* basis, given that the group throughput is equally shared by group members, can be set as the number of users forming group , so that group throughputs will be as much as possible proportional to group sizes.

Coefficients in Problem `Dynamic ABSF-PF` can be taken as a non-decreasing sequence of non-negative weights, so that past values of the throughput receive less or equal importance with respect to the prevision for next interval . For example, exponentially decaying coefficients or constant coefficients represent simple and widely adopted solutions for this kind of digital filtering problems.

Interestingly, Problem `Dynamic ABSF-PF` is simple to adapt also to cover the case in which the traffic of groups is not saturated. In such a case, the time window has to be smaller than the interval during which the set of receivers changes. In fact, during such interval all active receivers can be considered as saturated and the presented analysis holds.

8.5. Performance Evaluation

In this section, we evaluate our proposals—hereafter indicated as `Asymptotic ABSF-PF` and `Dynamic ABSF-PF` after the names of the optimization problems defined in Section 8.4.2—for several cellulars scenarios, including different mobility behaviors, group sizes and network area densities. In particular, we show that:

- the analytical model derived in Section 8.2 provides accurate results as validated against a packet-level simulator;
- the analytical model yields reliable results even when some of our assumptions do not hold (for instance, user groups are not collapsed in one spatial point) and under truly simulated user mobility models;
- our practical optimization solutions provide outstanding results in terms of throughput, transmission efficiency and user fairness, when compared to state-of-the-art approaches;

- `Asymptotic ABSF-PF` and `Dynamic ABSF-PF` outperform current standardized solutions when applied to optimize network operations in realistically evaluated scenarios, such as a dense-urban scenario covered with a heterogeneous cell deployment.

We have developed a Matlab event-driven packet simulator to study regular deployments, namely synthetic scenarios. We consider two radically different scenarios to show that our approach brings substantial benefits under very different operational conditions: a simple cellular network with base stations using the same transmission power (mW) and regularly spaced with Inter-site Distance (ISD) equal to m in a m m area, and the heterogeneous O2 deployment used for the example discussed in Section 8.3. Only downlink transmissions are taken into account in our simulation, with a MHz bandwidth and with the path loss model described in [111]. Power received by users, from the attached and the interfering eNB, follows an exponential distribution. We do not assume instead that the Signal-to-Interference-plus-Noise Ratio (SINR) follows an exponential distribution, like we did in Proposition 1, but we actually compute the SINR instantaneously as the ratio among the actual power received by the users.

Contrarily to what has been analyzed in Chapter 7, transmission queues are fully backlogged. Relay group can include up to users: whenever we will be dealing with the case of no cooperative eD2D communications. When active, base stations apply a weighted round robin policy to deliver the offered traffic to the relay groups, using the sizes of the groups as weights. We use the RWP mobility model [113] to move the center of gravity of each group within the simulated area, with a speed ranging from to m/s. For solving Problem `ABSF-PF` presented in Section 8.4.2, we run the ABSF optimizations every ms, with weights to achieve per-user fairness, , and . In all cases we use -bit long ABSF patterns cyclically repeated in the ms interval. Network simulations last s, whereas user channel conditions are evaluated on a subframe basis, e.g., each ms. For the case of the O2 deployment in London, we deal with a high-dense area in which the RWP mobility is developed through the streets on the map. All presented results are provided with confidence intervals.

To assess the performance of our practical solutions, we evaluate `Asymptotic ABSF-PF` and `Dynamic ABSF-PF` in terms of system throughput and fairness, the latter being measured by means of the well-known Jain's Fairness Index (JFI). For the sake of comparison, we also consider a solution without ABSF, namely `Legacy`, as well as another stochastic approach, using an optimization problem similar to `Dynamic ABSF-PF` that maximizes system throughput rather than user fairness, i.e., it optimizes the sum of group throughputs rather than the sum of logarithms. We refer to such policy as `Max Throughput`. We further benchmark our approach against two practical state of the art heuristics: `BSB`, proposed in [62], in which ABSF is used to target a max-min utility function; and `DRONEE` [21], in which relay groups (eD2D groups in our case) are formed dynamically to improve throughput. `BSB` and `DRONEE` were also fundamental components of TOMRAN, the coordination mechanism which we presented in Chapter 7. Lastly, we compare `Dynamic ABSF-PF` to standard randomized ABSF schemes with different ABSF application ratios.

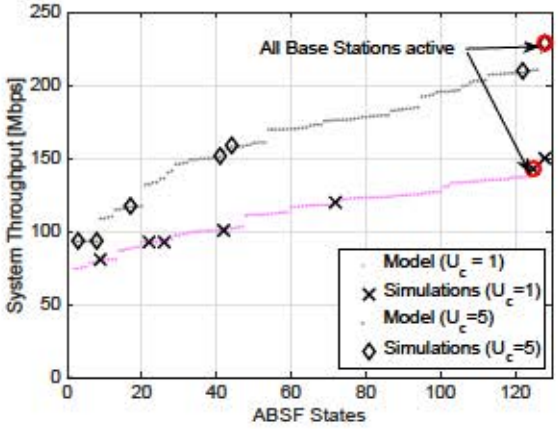


Figure 8.4: Model validation through exhaustive simulations.

8.5.1. Model Validation

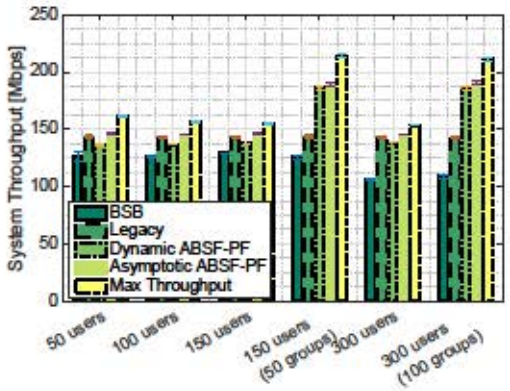
In order to evaluate and validate the analytical model presented in Section 8.2, in Fig. 8.4 we graphically provide a set of analytical results in terms of system throughputs . For the validation, we use static scenarios wherein the ABSF state and the position of nodes remain unchanged, so to compare the analytically derived throughputs (one for each state) with long-run averages per each state observed in simulations. Due to the huge computational effort required for simulating every possible ABSF state, only some significant ABSF states have been considered within the packet simulator. We mark with a red circle the ABSF state corresponding to all base stations simultaneously active, to point out the impact on the system throughput of no ABSF application. Notably, we observe that simulation results closely follow those provided through the analysis, properly validating the accuracy of our study with and without eD2D relay groups.

8.5.2. Performance in Homogeneous Deployments

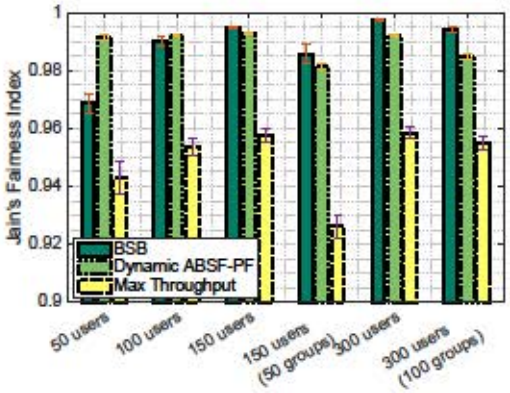
We next assess the performance of our practical solution for synthetic deployments of regularly spaced base stations using the same transmission power.

Figs. 8.5(a) and 8.5(b) present system throughputs and fairness levels achieved with different ABSF policies. Results are drawn for a few examples of user populations and numbers of relay groups. In the figures, the x-axis reports the number of considered users, and, when applicable, the number of relay groups. Each group consists of users, where is a uniform random variable drawn between and . Note that, figures report two cases with users, and two cases with users, i.e., with and without groups. Therefore, it is easy to observe both the impact of the user population size as well as of D2D communications.

In Fig. 8.5(a), the `Legacy` scheme shows an acceptable level of throughput even when compared to the `Max Throughput` scheme, but only when eD2D is not used. Our `Asymptotic ABSF-PF` and `Dynamic ABSF-PF` schemes provide reasonable results in absence of eD2D, outperforming the heuristic provided by `BSB`. `BSB`, which has not been though to handle a saturated scenario, heavily under performs if compared with the results in Fig. 7.4. Furthermore, our



(a) System throughput under different optimization policies.



(b) Jain's Fairness Index under different optimization policies.

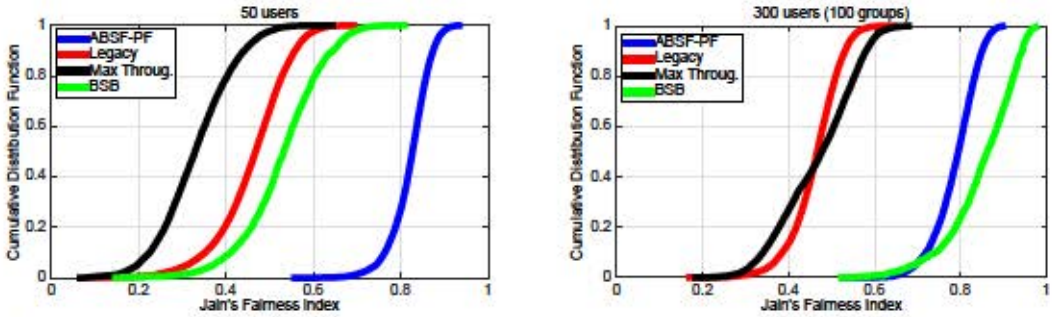
Figure 8.5: Performance Homogeneous Scenario

optimized mechanisms achieve similar throughputs as the `Legacy` scheme. Notably, in this homogeneous scenario, `Asymptotic ABSF-PF` obtains even better throughput than `Dynamic ABSF-PF`. However, both stochastic ABSF patterns do not help much in terms of throughput, unless eD2D is enabled. This confirms that ABSF, even when optimized, is not a suitable 5G solution on its own. Instead, in combination with eD2D, stochastic ABFS patterns make the difference. Moreover, `Legacy` and `BSB` schemes do not take advantage of eD2D relay groups, and their performance figures only slightly change with the number of users and groups. In contrast, `Max Throughput` leverages the transmission efficiency enhancements due to opportunistic eD2D relay and significantly boost throughputs. In all cases, `Max Throughput` represents the highest achievable network throughput. Therefore, Fig. 8.5(a) reveals the neat potential of eD2D and the fact that little gain can be expected by any scheme unless eD2D is jointly enforced.

Note that, as visible in Fig. 8.5(a), user density plays a very minor role in terms of system throughput. Conversely, density has a huge impact on fairness, as evaluated in Fig. 8.5(b), wherein we do not report results for `Legacy` and `Asymptotic ABSF-PF`, since those schemes are perfectly fair by definition in a completely homogeneous scenario like the one under analysis, at least on the long run.[3] In the figure, `Dynamic ABSF-PF` exhibits very powerful results when compared to `BSB` and `Max Throughput`. However, we need to remark that `BSB`, as the network becomes denser, shows better results in terms of fairness at the expenses of a very poor system throughput. Nonetheless, `Dynamic ABSF-PF` yields fairness levels very close to optimal fairness metrics in all cases (i.e., very close to). This confirms that stochastically-issued ABSF patterns and eD2D in combination are suitable for achieving high fairness while improving throughput.

For the same experiments, Figs. 8.6(a) and 8.6(b) report two examples of short fairness performance comparison. Every s, we computed the JFI of the average throughput of the users in

[3]Note that, in the homogeneous case, all users (groups) have the same spatial distribution. Thereby, `Legacy` as well as any ABSF strategy that does not change/adapt over time—as it happens with `Asymptotic ABSF-PF`—results in the same asymptotic throughput for all users (groups).



(a) Fairness distribution over samples collected every s. 50 users, no groups.

(b) Fairness distribution over samples collected every s. 300 users in groups of 3 members, on average.

Figure 8.6: Short term Fairness

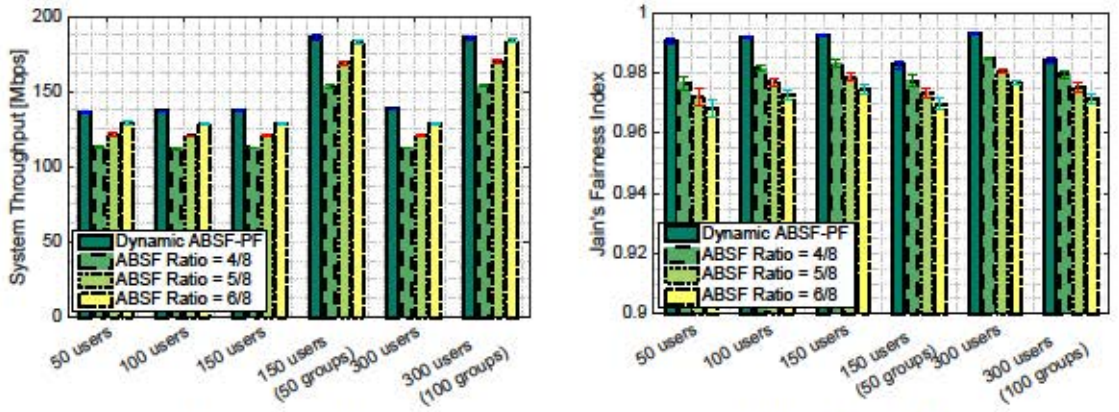
the system. The figures report distribution functions of the ontained JFI, from which it is evident that `Dynamic ABSF-PF` and `BSB` (the latter especially when the number of users is high) behave consistently fairer than `Legacy` and `Max Throughput` on short time scales. `ABSF-PF` shows very small variability in collected samples, which means that the optimization described in (8.16) promptly adapts to network changes caused by the mobility of users.

Next we compare `Dynamic ABSF-PF` to standard ABSF implementation with typical fixed application ratios, ranging from to (i.e., blanking from to of the subframes at random), as suggested in [58]. Here, blank subframes are randomly chosen by each base station independently, and the resulting patterns are automatically repeated to fill up the standard ABSF pattern of subframes. The 80-subframe long pattern is then repeated indefinitely [58]. Specifically, Fig. 8.7(a) compares system throughputs achieved with `Dynamic ABSF-PF` and when three fixed ABSF application ratios are applied to the system (with and without eD2D relay). The figure shows that standard ABSF schemes do not bring significant throughput improvements. Moreover, when fairness issues are considered, in Fig. 8.7(b), `Dynamic ABSF-PF` exhibits strong advantages with respect to fixed ABSF application ratios.

In summary, our stochastic ABSF scheme outperforms current standard solutions and offers a bargain trade-off between user fairness and spectral efficiency. `Asymptotic ABSF-PF` and `Dynamic ABSF-PF` provide extremely high fairness levels preserving reasonable throughput values, comparable to the maximum achievable when neglecting fairness issues. As final remark, exploiting eD2D relay communications is a notable advantage of our scheme.

8.5.3. Performance in a Heterogeneous Deployment

Since homogeneous base station deployments might bias our results, we next take into consideration a realistic scenario, focusing our attention on a particular use case: London city, as previously presented in Section 8.3 for the area reported in Fig. 8.2. We consider a total amount of users, as the user density for such area is users/km and the area considered is km [114]. Users are placed within the considered area following two distributions: one guides users mobility behavior along the streets while the other characterizes the static user po-



(a) System throughput.

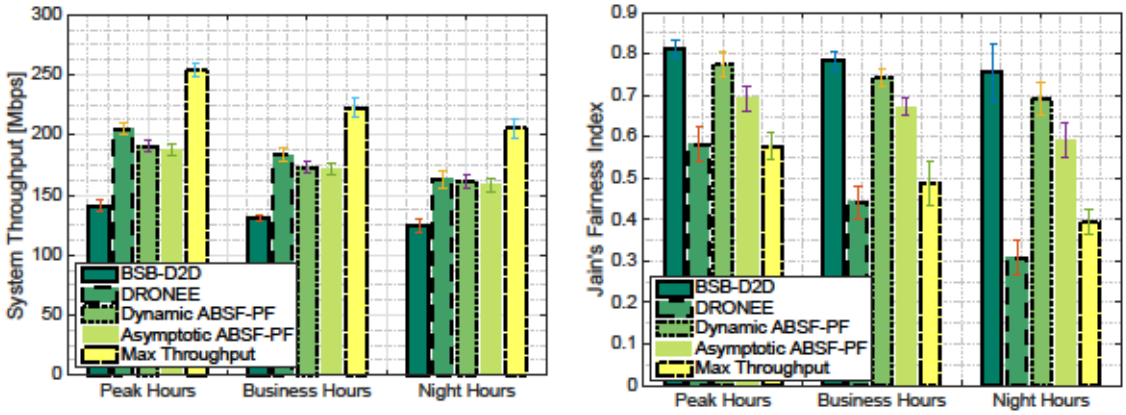
(b) Jain's Fairness Index.

Figure 8.7: Comparison with standard ABSF

sitions when they are within the buildings. Along our simulations, we vary the ratio between the averages of those two distributions to model different day-time periods. Mobile users follow a constrained RWP model: they select randomly a speed and a destination location within a valid street of the map, then they follow the shortest path and reach the new destination by following the streets of the map. When relay groups are in place, the group, e.g., its center of gravity, follows the mobility model rules on the streets while the users of the group are randomly placed around the center of gravity. Users within the buildings are statically allocated at random according to a uniform distribution.

Since we have shown in Sections 8.5.2 that eD2D relay is always beneficial, here we consider two benchmarking schemes in which eD2D groups are also present. In the first scheme, we allow the formation of eD2D groups also under `BSB`, i.e., we use our previously defined method Two-level Opportunistic Spectrum Management for 5G Radio Access Networks (TOMRAN), while in the second scheme we use the `DRONEE` mechanism defined in [21] to form clusters dynamically, under legacy base station operation (no ABSF). We refer to the resulting schemes as `TOMRAN` and `DRONEE`, respectively. We have carried out different simulations to evaluate the realistic deployment in different operational timeframes. For each timeframe, we properly model the ratio between the distribution of users moving along the streets and the users staying within the buildings in the following way: *Peak Hours*, during lunch time, , *Business Hours*, during morning and afternoon, , *Night Hours*, [114].

In Fig. 8.8(a), we show the system throughput expressed for different schemes. We observe that the system throughput increases during peak hours, as most of the people are moving outside and, thus, exploiting opportunistic eD2D connections brings an additional gain. `Asymptotic ABSF-PF` and `Dynamc ABSF-PF` perform quite well, showing similar throughput figures as `DRONEE` and significantly worse throughput results only if compared with `Max Throughput`. Note that `Asymptotic ABSF-PF` obtains only slightly lower throughput than `Dynamic ABSF-PF`, although it results in much lower fairness. More in detail, due to the heterogeneity of the realistic scenario, user fairness is significantly impaired compared to results obtained



(a) System throughput in London deployment.

(b) Jain's Fairness Index in London deployment.

Figure 8.8: Performance Heterogeneous Scenario.

for a homogeneous deployment, as shown in Fig. 8.8(b). However, `Dynamic ABSF-PF` unveils the great potentials of properly applying a dynamic ABSF scheme, outperforming not only the asymptotic optimization scheme, but also `Max Throughput` and `DRONEE` solutions in terms of JFI by about and , respectively. Moreover, the fairness achieved with `TOMRAN` is comparable with or better than our proposal's one, though it provides much less throughput.

In conclusion, our results illustrate how `Dynamic ABSF-PF` is intrinsically better than standard approaches and pure (static) ABSF optimizations targeting throughput or low interference. In particular, although we have shown that ABSF alone is not able to boost throughput, `Dynamic ABSF-PF` manages to handle the throughput enhancements achievable with eD2D relay while achieving very high fairness levels. However, in case of homogeneous scenarios, the importance of dynamic optimization becomes lower, and an asymptotic and less complex optimization approach can be used instead.

Chapter 9

Convergence to Multi-resource Fairness Under End-to-end Window Control

Up to now, we focused only on the last wireless hop of the Radio Access Network (RAN). Nevertheless, in the next generation of wireless networks, flows may have to traverse several resources before reaching destination. Centralized-Radio Access Network (C-RAN) have been indeed proposed as a possible solution to reduce the complexity of the densification of access networks. A C-RAN consists in multiple radio PoAs performing only signal transmission/reception, a fronthaul/backhaul connection, a transport network and a central data center. The data center manages a pool of virtual base stations, each representing one or more actual base stations, which performs the majority of the base-band processing. In the wired Internet the resources to traverse are links and all flows have the same requirement for each bit/s of rate. In C-RANs, though, the resources traversed by the flows are wireless links and computational resources and requirements may be heterogeneous. For example, the amount of spectrum consumed for each bit/s depends on the flow's radio conditions. In the same way, the amount of CPU/RAM required for encoding/decoding depends on the number of information bits transmitted per each OFDM symbol and, somehow, on radio conditions.

Resource sharing by flows with homogeneous requirements has been widely studied over many years. Of particular interest for the present chapter is the observation made some 30 years ago that end-to-end max-min fairness is realized by implementing fair queuing in router queues and performing window-based flow control [69]. Our main objective here is to derive the equivalent result in the case of heterogeneous resource requirements. Explicitly, we show that imposing local fairness at each resource, coupled with end-to-end flow control, results in a desirable generalization of max-min fairness called Bottleneck Max Fairness (BMF) [9].

Proving this result is hard due to the complex dynamics of per-flow resource queue backlogs. For homogeneous requirements, the proof that backlogs eventually converge and rates stabilize at their max-min fair shares was the culmination of several years doctoral thesis work by Hahne [72]. Chrysos and Katevenis have since derived a somewhat simpler proof, thanks to their use of a fluid

model, but this is still highly non-trivial and again confined to hop-by-hop window control [73].

We apply the same fluid model as Chrysos and Katevenis to prove convergence to BMF for flows with heterogeneous requirements in a network limited to two resources. It is considerably harder to account for heterogeneous requirements because the water filling characterization of max-min fairness used in [71] and [73] does not generalize to BMF. In addition, we have used water filling with insights gained from the analysis of BMF to derive an original proof of convergence for a general network with homogeneous requirements under end-to-end window control.

After the mathematical analysis enabled by the fluid model adopted, we evaluate the approximation introduced to the theoretical objective of multi-resource sharing in a practical packet-based scenario. Our goal is indeed understanding how a more realistic assumption on the arrival traffic deviates the real system performance from its ideal behavior. We simulate a network where resources implement Start-Time Fair Queuing (SFQ)[1] [24] and investigate the impact on convergence times of window size, the number of competing flows and their particular requirements. Finally we apply local fairness at each resource in the scenario which motivated this work, i.e., a C-RAN-like access network. In the scenario we also consider D2D communications. D2D, indeed, proved to be a fundamental tool to improve legacy cellular access network and we envision its use also in the new generation of cellular network. Nevertheless, when D2D is used for relay purposes, D2D implies that flows have to traverse also an additional resource, i.e., the cluster-head, and multi-resource fairness among flows becomes of astounding importance.

In the next section, we argue the need to perform fair resource scheduling for heterogeneous requirements and discuss properties of the resulting BMF allocation. Section 9.2 introduces the dynamical system describing the evolution of backlogs in the fluid model limit. The main convergence results for this dynamical system are given in Section 9.3. Section 9.4 presents simulation results that illustrate deviations from the fluid model ideal when accounting for finite sized packets, while Section 9.5 shows their application to a C-RAN environment.

9.1. Multi-resource Sharing

We discuss why scheduling is required for fair multi-resource sharing before recalling the desirable properties of BMF. The main symbols used are listed in Table 9.1.

9.1.1. Need for Scheduling

Consider a network with resources shared by flows. Resource has capacity of units per second where units are resource dependent. Flow requires units of resource to process

[1] SFQ is a scheduling algorithm which assigns almost ideally equal shares to each flow at any resource, even in a packet system

	number of resources
	number of flows
	capacity of resource
	requirement of one unit of flow at resource
	normalized requirement
	resource visited by flow prior to visiting resource
	rate at which flow leaves resource
	rate allocated to flow
	fair share at resource
	backlog of flow at resource
	backlog indicator, 1 if , 0 otherwise
	end-to-end window of flow in bits
	end-to-end window in packets

Table 9.1: Summary of notation.

each bit. The rate in bit/s, , allocated to flow must satisfy the capacity constraints

(9.1)

for . is the amount of resource used per second by flow .

In C-RAN, for example, the envisaged resource use is the following:

- central data center: the resource consists in the CPU/RAM used for coding/encoding. is measured in cycle/s and is the number of cycles needed to process one bit of flow ; [8] suggests that the higher is the number of informative bits contained in each symbol, i.e., the higher is the complexity of the MCS used, the higher is the number of cycles required for encoding/decoding.

- fronthaul/backhaul wired network: is measured in bit/s; if flow uses link and otherwise;

- LTE wireless link: is measured in symbols/s; is the fractional number of slots needed to transmit each bit of flow accounting for the flow's radio conditions; the requirement can be more than 20 times smaller for a user close to the antenna than for a user at the cell edge;

- D2D relay link: if out-band D2D is used to relay the traffic, is 1, and can be approximated as the fraction of time needed to transmit each bit from the cluster-head to the corresponding cluster-member.

Scheduling is generally absent in the wired Internet and bandwidth sharing is realized by means of congestion control protocols like TCP that react to drop signals received from First-In First-Out (FIFO) buffers. Sharing is generally fair enough if users implement the same protocol

[76] though it has often been noted that fair queuing implemented in router queues would provide more robust control, e.g., [77], [78].

For a wireless link, where requirements are highly variable, it is generally considered preferable to aim for equal resource shares, , rather than equal bit rates. This is broadly what the proportional fair scheduler achieves [79], as implemented in 3G and 4G cellular networks. That the IEEE scheduler tends to realize max-min fair *bit rates* was recognized as a performance anomaly [80]. This anomaly is mitigated in more recent WiFi versions, like , where packet aggregation indeed tends to equalize the . In C-RAN, the dynamic provision of compute capacity implies CPU may become a temporary bottleneck and the way it is shared is therefore an issue. Simple FIFO queuing coupled with end-to-end congestion control would lead to approximately equal flow bit rates, as in a wired Internet. However, if requirements differ significantly, max-min fair rates would produce the same "performance anomaly" as in .

In this work we adopt the position that flows using a single resource considered in isolation should receive max-min fair resource shares. Let be the incoming bit rate of flow at resource . The scheduler allocates a fair share to any flow such that , and allocates to the others, where fair share is determined by the capacity constraint (9.1).

9.1.2. Bottleneck Max Fairness

Ghodsi and co-authors introduced the problem of multi-resource sharing in compute clusters [74] and extended their analysis to networks [75]. They advocate so-called Dominant Resource Fairness (DRF). In networking applications, DRF requires schedulers at each resource to implement weighted max-min fairness with the same flow weight applied at each resource determined from the dominant relative resource requirement, . This choice is motivated by a requirement that the allocation be *strategyproof*: flows should not be able to gain a greater bit rate by falsely stating their requirements.

The plausibility of designing and implementing such a gaming strategy in a context of dynamic demand in a network setting is highly debatable, however. BMF is an alternative allocation technique that sacrifices strict strategyproofness in order to achieve a better efficiency–performance tradeoff.

Like max-min fairness, BMF is defined for a fluid model where packet size is infinitesimally small and resource capacity is perfectly divisible among flows. The allocation is such that resource sharing is Pareto efficient (i.e., all capacity is used if possible) and every flow receives the maximum allocation at some resource that is fully used. This may be recognized as one of the definitions of max-min fair resource sharing, e.g., [115]. The significant difference here derives from the heterogeneous requirements in capacity constraints (9.1).

It was shown in [9] that the BMF allocation always exists and that it has the so-called *single-bottleneck fairness* property: if the network has a unique bottleneck , the allocation is such that for all flows . Such allocation would also be the one that Proportional Fair (PF)

would achieve in the same case. That this property is not shared by DRF largely explains its inferior throughput performance.

A significant advantage of BMF in networking applications is that the allocation can be realized simply by implementing weighted fair queuing independently at each resource with weights for flow at resource equal to . The main objective of this work is to justify this statement.

9.1.3. Scheduling and Window-based Flow Control

We suppose flow maintains a fixed volume of unacknowledged data and every resource realizes weighted max-min fair sharing with weights . Assume the network attains a steady state with constant flow rates , constant queue backlogs and constant round trip times.

Proposition 1. *Suppose flows implement a large enough fixed window and resources realize weighted fair queuing. If the network attains a steady state, the realized flow rates are bottleneck max fair.*

Proof: The proof is immediate as the window can be made large enough that every flow has at least one bottleneck resource (i.e., it has a backlog and the resource is therefore fully used) while the scheduler ensures its share of that resource is maximal. These, with Pareto efficiency, are the conditions that define BMF.

This proposition also applies to max-min fairness as a special case of BMF and its equivalent was stated by Hahne [71]. All the difficulty in proving the controls yield BMF is in proving the system does in fact converge to a steady state.

9.2. A Dynamical System

We present the dynamical system governing the evolution of the resource queue backlogs and flow rates under a fluid model with zero propagation times. We assume resources are consumed successively in the order defined by a flow-specific route: designates the resource visited by flow prior to its visit to resource . In this and the next section, to simplify the formulas, we use normalized requirements .

9.2.1. Persistent Binary System States

Let denote the rate at which flow is served by resource at time . For brevity, we generally omit the explicit dependence on time in this and other variables. Flow rates depend on the backlogs at each queue or, more succinctly, on the backlog status indicators ,

In periods where the are constant, rates are also constant and satisfy the following equations

where

$$\frac{\mathbb{1}}{\mathbb{1}}$$

These equations can be rewritten:

$$\frac{}{} \tag{9.2}$$

They express the result of per-resource weighted max-min fair schedulers, as described at the end of Section 9.1.1.

To avoid unhelpful complications, we suppose the are such that linear equations (9.2) are independent and therefore yield a unique set of 's for each binary state vector . For many such vectors, the computed 's will not in fact be feasible (e.g., they might be negative). Vectors that do yield a set of feasible rates constitute the space of valid *persistent binary states.*

9.2.2. Evolution Between Persistent States

The system evolves between persistent states as follows. All backlogs for which are decreasing in time. Let the queue of flow at resource be the one to empty first. At this instant the system enters a new state where and for the other queues.

If is a persistent state (i.e., there is a feasible solution to the new instance of equations (9.2)), the system will enter a new phase with a new set of rates which persist until a new queue empties. If is not persistent, some of the non-backlogged queues will immediately become backlogged because the flow's incoming rate exceeds its weighted fair share. In case is not persistent, the rates satisfy the following equations:

(9.3)

where the fair share satisfies

$$\frac{\mathbb{1}}{\mathbb{1}}$$

The new persistent state is where if and , and otherwise. Notice that is indeed a persistent state since the satisfy (9.2) with replaced by .

9.2.3. Convergence

The network evolves between persistent states until it enters one in which the service rates at all resources are the same for each flow. If this occurs, the rates in question are BMF by Proposition 1.

The equations defining the dynamical system can be rapidly solved numerically allowing us to explore convergence over a wide range of parameter values (e.g., 1 million random choices). Convergence indeed always occurs in all our experiments. The graph defined by valid transitions between persistent states is acyclic. There are cases where the BMF allocation is not unique [9]. In such cases, numerical experiments show the system converges to one or another of the possible allocations, depending on the assumed initial backlogs .

Unfortunately, it proves very difficult to analytically prove convergence in general. The next section proves convergence for some significant special cases.

9.3. Proof of Convergence

We first prove convergence to BMF for networks of 2 resources consumed successively before discussing the difficulty of extending this result to more resources. We then prove convergence for resources in the special case where BMF reduces to max-min fairness, i.e., in case of homogeneous requirements.

9.3.1. BMF for 2 resources

We consider a system with 2 resources and flows. The capacity of all resources is normalized to 1. To avoid some tedious qualifications, we suppose here that and the ratios are distinct.

Inspecting the results of numerous simulations, we observed that the evolution of backlogs is such that one queue in particular is never empty in any persistent state. The resource in question is and the flow is . For convenience and without loss of generality we therefore renumber the resources such that

— — (9.4)

and the flows such that

— — — (9.5)

The queue that never empties is then that of flow 1 at resource 1. The following three lemmas allow us to affirm convergence in Theorem 8. For a better visualization of the main result of this section, Lemmas 3 and 5 are proved in Section 9.6 while the proof of Lemma 4 is symmetrical to that of Lemma 3 and is therefore omitted.

Lemma 3. *Flow 1 always stabilizes to a backlog only at resource 1. Given flows 1 to are backlogged only at resource 1, for , a sufficient condition that flow also stabilizes to a backlog only at 1 is*

(9.6)

Lemma 4. *A sufficient condition for flow to stabilize to a backlog only at resource 2 is . Given flows to , for , are backlogged only at resource 2, a sufficient condition that flow also stabilizes to a backlog only at 2 is*

(9.7)

Lemma 5. *Given the first flows for have converged to a backlog only at resource 1 and the last flows for have converged to a backlog only at resource 2, either flow will converge to a resource 1 backlog, or flow will converge to a resource 2 backlog.*

Lemma 5 shows that the system will eventually converge to a state where the first flows stabilize with a backlog only at resource 1 while the remainder stabilize with a backlog only at resource 2 for some , . The order of convergence depends on initial backlogs and the various flow window sizes. The lemma implies that convergence will occur at least as fast as if first flow 1 stabilizes, then either flow 2 or flow , and so on, proceeding from either the lowest or the highest numbered remaining flow. In this order, the last flow to stabilize is either flow or flow . The value of characterizes the BMF allocation and is specified in the following theorem.

Theorem 8. *When each resource locally realizes weighted max-min fair sharing with respective weights , flow rates converge to the BMF allocation. When resources and flows are labelled such that* (9.4) *and* (9.5) *hold, flows 1 to are backlogged only at resource 1 and the remainder only at resource 2, where is the index such that*

(9.8)

The BMF allocations are , for , and , for , where and are the fair shares

(9.9)

(9.10)

Proof. We know from Lemmas 3, 4 and 5 that the backlogs of all flows eventually stabilize and Proposition 1 shows the resulting allocation is BMF.

Condition (9.6) may be written where and are the numerator and denominator of the right hand side, respectively. It may readily be verified that for . This condition therefore ensures by Lemma 3 that the first flows stabilize with a backlog only at resource 1.

If now , which is the second inequality in (9.8), we can similarly show that for . Condition (9.7) is thus satisfied for flows to completing the proof that (9.8) characterizes the stabilized backlogs.

Finally, fair shares (9.9) and (9.10) follow from the equations

—

—

9.3.2. More than 2 Resources

While all our simulations of the dynamical fluid system confirm convergence, at the time of writing, we have not been able to prove this analytically. We observe in the numerical results that, in every case, the backlog of one flow at one resource is always decreasing or empty. Unfortunately, this queue is not identified by a simple generalization of the characterization discovered for 2-resource networks. It depends on the routing but the same BMF allocation results from all possible routings. The next section provides an original proof of convergence for a network of any size in the special case of max-min fairness.

9.3.3. Max-min Fairness

The resources considered here are wired network links. Consider a network of links of capacities , for , shared with max-min fairness by flows. Flow has requirement when it uses link and otherwise. Flow maintains a window of unacknowledged data. The proof of convergence of the corresponding dynamical system derives from the well-known water filling definition of max-min fairness, as used previously by Hahne [71] and Chrysos and Katevenis [73] for hop-by-hop window control.

Let be the set of order 1 bottlenecks defined by

—

contains the links that are simultaneously saturated first in the water filling procedure. Let be the set of flows such that for , i.e., the flows that use at least one of the order

1 bottlenecks. Now define recursively bottleneck sets of order and corresponding flow sets by

where is the weighted max-min fair rate for flow and

for or

Links are saturated in step of the water filling procedure.

The weighted max-min rates are given recursively by

(9.11)

for , and

(9.12)

for and .

Theorem 9. *When flows have homogeneous requirements and are controlled by an end-to-end window, rates converge from any initial state to the max-min fair rates.*

Proof. The proof is by induction. We first prove that buffers at links for flows are drained at rate and their backlog is either increasing or stable.

If the backlog of flow at some link is less than , flow must be backlogged at some other link. Denote by the first backlogged link in the path of preceding . The rate of flow leaving link satisfies

Rate is the rate into and, by the definition of , . This is true for all flows using whose backlog at that link is less than (including those that have no backlog because link follows some other link in in the flow path).

Let be the service rate of flow at link . Note that is the max-min fair rate realized locally by flows using when all flows are backlogged at . For any flow backlogged at , cannot be less than . On the other hand, any non-backlogged flow must be served at its input rate and we have just shown that this is at least equal to . Clearly, the only allocations that satisfy these conditions and the capacity constraint, , are the given by (9.11).

Increasing queues (i.e., where) must stabilize before some time after which no link outside has a backlog for flows . When contains more than 1 link, the stable backlogs of flow can be any partition of the window .

Now suppose flows have converged to their fair rates and that after some epoch their windows are entirely contained in the buffer of some link in . We need to prove

flows will then similarly converge after some epoch .

Consider . If the queue at some link is less than , flow must be backlogged at some other link. Denote by the first link in the path of preceding to have a flow backlog. The rate of flow leaving link satisfies

As for , this implies the output rate is at least that given by (9.12) for these flows. As any other flow in with a full backlog cannot receive a lower rate, we conclude they must all receive the same rate given by (9.12).

The backlog is either increasing or stable. If all queues are already stable, the induction hypothesis is satisfied and . If not, all increasing queues will have stabilized at some later epoch after which no link other than has a backlog for flows .

Note that the special case of BMF where requirements are either equal to some flow dependent value or 0 is equivalent to max-min fairness for the resource shares . Theorem 9 therefore has the corollary that this special case indeed converges to the BMF allocations.

9.4. Packet System Behavior

Simulation is used to evaluate the convergence behavior of packetized flows. Results show that flows indeed converge to the fluid BMF rates as long as the window in packets is large enough.

9.4.1. Packet Model

We simulate a network of 2 unit capacity resources where all flows use both resource. Packets are of constant size bits and are distinguished by flow requirements . Propagation times are zero. The source of each flow maintains unacknowledged packets in the backlog of either resource (in the fluid model notation). As packet size and window are fixed, it is as if packets circulate from one resource to the other and, in results below, we measure time in round trip times (RTTs), the variable time between successive service completions of the "same" packet.

Each resource implements (SFQ) [24]. The start time tag of the packet of flow to arrive at resource is computed recursively on its arrival epoch by,

(9.13)

where is the start time of the last packet to have begun service at . Packets are served in increasing order of start time tags. In terms of the algorithm described in [24], it is as if packets

have length .

SFQ only approximates weighted max-min fairness and resource sharing realized by the packetized flows differs from the fluid ideal. In particular, the BMF rates may not be attained if the window is too small. The following proposition gives a lower bound on the required window size.

Proposition 2. *A sufficient condition for the flows to attain the BMF allocation in the considered 2 resource network is that the end-to-end window in packets satisfies,*

— (9.14)

for all flows and all resources .

Proof. Th. 1 of [24] shows that the difference in the amount of service received by two continuously backlogged flows in a given interval is bounded. Reinterpreting the parameters of the bound in terms of the present network, we have, for two flows and , backlogged throughout interval

,

where and are the number of flow and packets served in the interval. In order that the flows be backlogged, it is necessary that be large enough to absorb the fluctuations. The requirement for flow is for at least bits in the backlog to effectively satisfy the momentary excess service rate. This translates to a window — since each packet has an effective "length" of bits. Repeating this reasoning for each possible couple of backlogged flows and each resource yields (9.14).

9.4.2. Tending to the Fluid Limit

The fluid model occurs in the limit where the packet size tends to zero while the window in volume of data remains fixed. The backlog and throughput results derived for the fluid model in fact predict the performance of the packetized system for quite large packets.

Fig.9.1 shows the evolution of the share of in the resource 1 backlog as a function of the number of RTTs experienced by flow 2. The figure relates to a particular instance where 2 flows share the 2 resources with respective requirements , , , . The window of each flow is initially evenly split between both resources.

The red lines, for , almost coincide with the results of the fluid model (black lines). For larger packets, with , the backlogs follow the fluid trends but naturally exhibit greater variability.

In Fig.9.2 we show the impact of on the evolution of flow rates at resource 2. The rate received by flow at resource in the packet model is defined as

—— — (9.15)

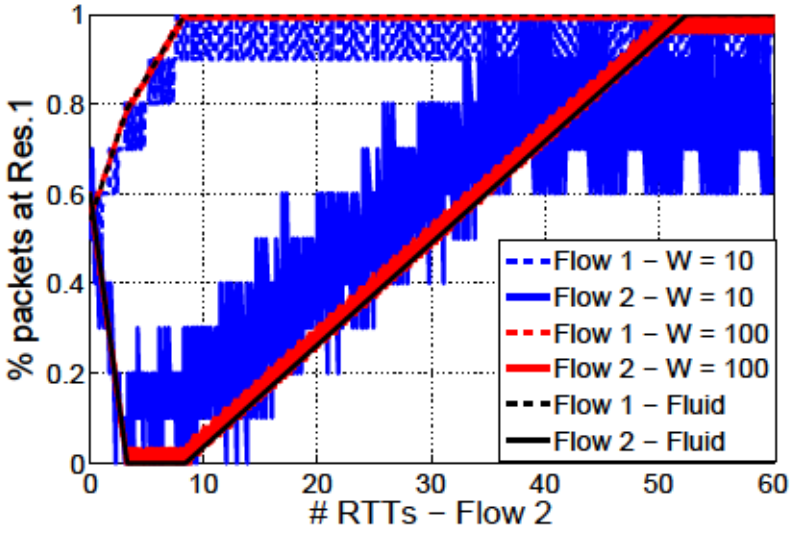


Figure 9.1: Example of Backlog evolution.

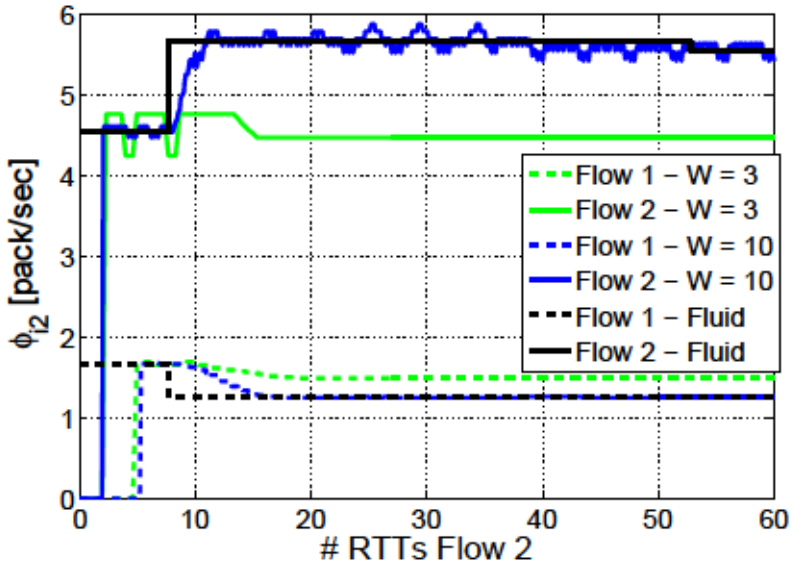


Figure 9.2: Example of unfair convergence.

where $P_{ij}(t)$ is the number of packets of flow i served before time t at resource j and T_{kj} is the RTT of packet k at resource j. Variable $\hat{\phi}_{ij}(t)$ is an average of the flow i rate leaving resource j over the last $W^{(p)}$ packets.

Convergence to BMF rates occurs if $W^{(p)}$ is larger than the bound of Proposition 2. In the present example, the rate of flow 2 for $W^{(p)} = 3$ converges to a value less than the BMF allocation. On the other hand, for $W^{(p)} = 10$ and greater, convergence to the fluid limit rates occurs within the first few RTTs for resource 1 (not shown). Convergence takes 50 RTTs however for rates at resource 2.

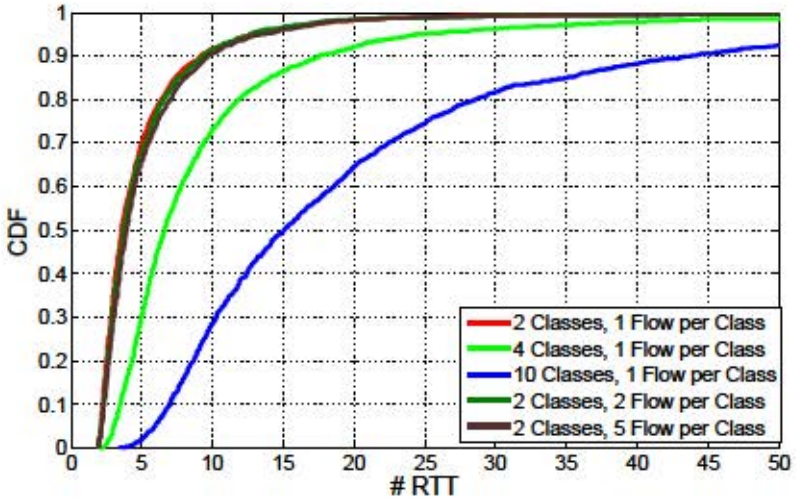


Figure 9.3: CDF of convergence time to BMF rates.

9.4.3. Speed of Convergence

The time to converge to the BMF rates depends on the requirement parameters. We have explored this dependence by simulating the 2 resource network with many different random requirements. Flows of the same "class" have the same requirements drawn uniformly at random between 0 and 1. There is a variable number of flows per class. We set for this experiment. The initial assignment of the packets of each flow at each resource is also set at random.

Figure 9.3 plots the cumulative distribution function of the time to convergence for 5000 realizations. The rates are deemed to converge when all measured rates, , are simultaneously within 5% of the calculated BMF rates. Time to convergence for each flow is measured in its own RTTs and the overall convergence time is the maximum of these counts.

The figure shows that convergence time can be quite long and increases statistically with the number of flows. Slow convergence occurs when a flow backlog drains from one resource to the other at a very slow rate since the fair share at each is momentarily very similar. The occurrence of such an event is more likely as the number of flows with different requirements is greater. This explains why the convergence time in this experiment is statistically longer for the cases with more classes. Note, however, that even when convergence to precise BMF rates may be relatively long, it is very rare (% of cases) that the rate attained by any flow after 20 RTTs differs by more than 20% from its BMF rate.

We have also investigated the impact on convergence time of the packet window size . The convergence time in RTTs has roughly the same CDF if is somewhat greater than the bound in Proposition 2. This CDF coincides with that of the fluid model evaluated using the same random choice of requirements and initial backlogs.

9.5. C-RAN Multi-resource Sharing

In this section, we apply the multi-resource concepts presented above to the uplink of a D2D-assisted C-RAN (even though the same approach may be used also in downlink). In the uplink of a D2D-assisted C-RAN, each flow has to traverse several resources before reaching the core network. The possible cases are two:

- the flow belongs to a relay node, i.e., to a user which is directly connected to an eNB. In this case, the flow first traverses the uplink channel towards the attached eNB, then the mm-waves (or optical) backhaul connecting the eNB to the datacenter and, finally, gets decoded at the datacenter. For the uplink channel and the backhaul, the flow contends the resources (RBs and airtime) with all the users attached to the same eNB, while at the datacenter, it contends the computational resources with all the users attached to one of the eNBs managed by the datacenter;

- the flow belongs to a Group Member (GM), i.e., to a user which is connected to an eNB only via a relay node. In this particular case, we envision the use of out-band D2D communications and, hence, the use of the bandwidth. The flow belonging to a GM first contends the resources with all the users belonging to the same relay group. Then, after arriving at the relay node, the flow traverses the same resources as the ones traversed by the relay node's flow.

In a C-RAN, the performance bottlenecks of a flow may consist in the bandwidth, in case the relay node is managing a large group, the uplink channel or also the datacenter (depending on its computational capacity and the pool of eNB it has to manage). In the following, we assume instead the backhaul links have enough capacity to accommodate the peak of traffic which can be sent in uplink to an eNB. Therefore, we do not consider the backhaul links as a possible performance bottleneck.

In this context, our main goal is to achieve multi-resource sharing with the least possible changes to the existing protocols. Therefore, we use state-of-the-art schedulers which perform local fairness and we assume that the datacenter acknowledge to the users the correct reception of each block of traffic in order to apply window control. This procedure is already in place in LTE, since eNBs have to acknowledge the correct reception of each block of traffic within after user data transmission.

In order to implement local fairness at each resource, we use the following schedulers:

- In bandwidth we use standard . As we previously mentioned, equalizes the share of airtime assigned to each user on the WiFi bandwidth via frame aggregation. The number of consecutive packets a user can send each time it successfully access the channel is proportional to its channel quality, i.e., inversely proportional to packet time transmission.

- In the uplink cellular channel, flows are scheduled following Round Robin (RR). Each , the eNB uses SC-FDMA to schedule users over its bandwidth, trying to assign the same number of RBs to each user. In case this is not possible, in the following , the eNB schedules first the users which quota of previously assigned RBs is the smaller. The RR scheduler takes into consideration that each relay node is transmitting the traffic of several other users. Hence, each flow, independently from the fact that it arrives directly to the eNB or via the relay node, receives its own fair share.

- At the datacenter, flows are decoded in parallel. Computational power, also in this case, is shared equally among flows.

Please note that the scheduling algorithms used at the different resources react automatically to flows which are not saturated. When a flow does not have the amount of traffic to fill its fair share, indeed, the exceeding quota of resources is shared among the backlogged flows. This is important because, at convergence, the window of packets of each flow is positioned at the corresponding bottlenecks, while the rest of the resources will be seeing as arrivals only the traffic that the bottlenecks are able to serve, i.e., less than the fair share.

9.5.1. Performance Evaluation

In the following, we evaluate by simulations the achieved performances of the proposed multi-resource sharing mechanism over a realistic scenario.

As simulation set-up, we choose again the London scenario presented in Section 8.3. The scenario is heterogeneous, with small and macro cells (see Fig. 8.2 for more details). Users are attached to the eNBs depending on the power sensed in downlink from the base stations, but they all transmit in uplink with the same power ( ). Both the WiFi bandwidth and the cellular bandwidth are MHz, while the path loss model is again the one described in [111]. Groups can be formed only among users far at most meters to each others. We analyze the system in a period where the users do not move and where the SNR sensed at the eNBs and at the relay nodes is fixed. In the period analyzed, therefore, also groups do not change and they are therefore an input of our problem. Given the assumptions proposed, the requirements of each flow remain constant over time, and it is possible to compute the requirement matrix .

In this section we aim at showing two different results. First, we show the performance of a C-RAN which implements the same schedulers presented in Section 9.5 at each resource, without the window control algorithm. Second, we show the performance of our sharing algorithm compared to an ideal proportional fair mechanism. For each simulation, we retrieve the requirement matrix and we obtain via optimization the set of throughputs which maximize , respecting the capacity constraints.

We built a simulator in MATLAB which simulates the MAC of for the groups, the RR resource scheduling of Long Term Evolution (LTE) and the parallel execution of the tasks in the datacenter. In the simulator, whenever a packet is served at the datacenter, a new

packet is allowed to be transmitted at the source of the served flow. In the simulator: the SNR sensed at the relay node dictates the number of bits Hz transmitted over the WiFi groups. The mapping SNR/channel quality is performed following [116]; the SNR sensed at the eNB dictates the number of bits symbol transmitted over the uplink channel. The mapping SNR MCS is performed following [58]; the SNR sensed at the eNB also dictates how difficult it is deconding the received transmission. The number of bits iteration for the coding is given by [8].

Interestingly, the users with large requirements for the wireless link, i.e., the users with low SNR, typically have a small requirement at the datacenter, since their transmissions carry few information bits. On the contrary, the users with very good channel quality, i.e., with small wireless requirements, present very large requirements at the datacenter. When flows have mixed characteristics, multi-resource sharing becomes fundamental to infer end-to-end fairness.

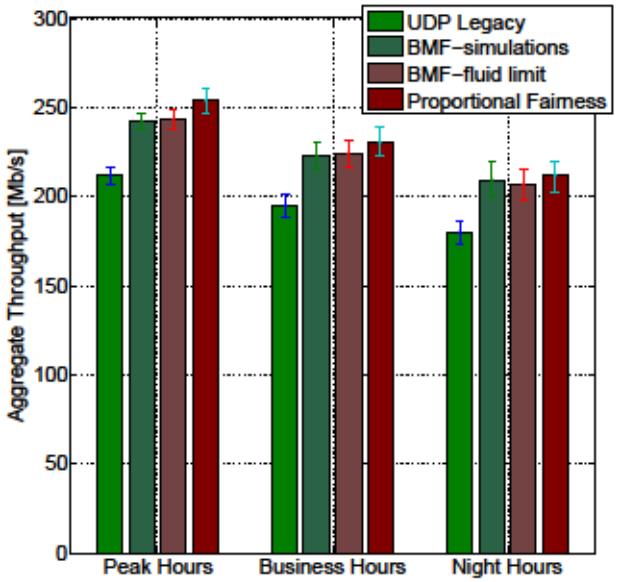
In Fig. 9.4, we show the London scenario results when the users in the system are users. As in Chapter 8, we evaluate the system under three different conditions: *Peak Hours*, during lunch time, where the of the users is outside the buildings, *Business Hours*, during morning and afternoon, where the of the users is outside the buildings, *Night Hours*, with only of the users outside [114]. Also in these simulations, only the users outside the building are allowed to form *static* groups. In this way we can also factor in the uplink case the effects of D2D, since different simulations present different densities of groups presence. We evaluate the performance of the C-RAN in terms of aggregate throughput and utility ( ).

We first compare the performance of BMF obtained through simulations and using the backlog evolution at the fluid limit which we explained in Section 9.2. As it is possible to see from Fig. 9.4, the fluid limit modeling depicts quite accurately the behavior of the real system, since aggregate throughput and utility are very close in both cases. Furthermore, even if the cases under analysis are much larger than resources, all the simulations performed converge to a particular backlog state, which is BMF.

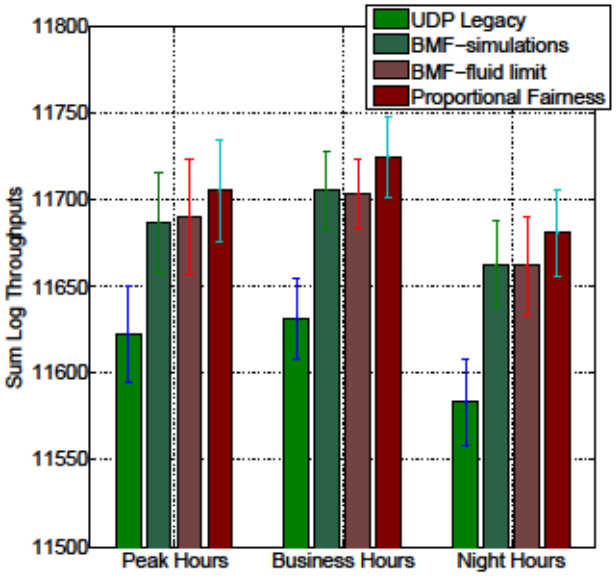
Then we compare the performance of BMF with a system implementing the fair schedulers at the resources, but not the window control mechanism. We name such a solution "*UDP Legacy*". It is clear from Fig. 9.4 that the use of a window control is fundamental to improve the performance of C-RANs, both in terms of aggregate throughput and in terms of utility, since it avoids unnecessary waste of resources by the flows which are bottlenecked somewhere else in the network.

Finally, we compare the results we would achieve if we were able to perform perfect proportional fairness. In a distributed environment, implementing proportional fairness is computationally expensive, since information about backlogs should be exchanged among resources [9]. Nevertheless, proportional fairness represents a valid benchmark to see how far BMF performs from the optimum. As it is possible to see from Fig. 9.4, BMF is close to the optimum both in terms of aggregate throughput and utility.

All in all, BMF proves to be a distributed algorithm which can be implemented very easily with state-of-the-art protocols and approximates quite closely, even in a realistic scenario, the op-



(a) C-RAN Aggregate Throughput.



(b) C-RAN Utility ( ).

Figure 9.4: C-RAN performance under different mechanisms

timal proportional fairness algorithm. BMF is a natural extension of the optimization techniques presented in this thesis, since it applies one step further than the last wireless link.

9.6. Proof of Lemmas 3 and 5

The following is used in the proofs of Lemmas 3 – 5.

Lemma 6. *Inequality* (9.4) *implies* — .

Proof. Since the harmonic mean of a set of numbers is always smaller than the arithmetic mean,

— — — —

Thus,

—

and the lemma follows since the are all positive.

Lemma 3 – a Sufficient Condition

Proof. When , the denominator in (9.6) is zero and the condition reads since the right hand side must be negative. With this interpretation we prove that (9.6) is a sufficient condition for flow to be backlogged for . Since by Lemma 6, sufficiency is enough to prove the first statement.

We only need to consider evolutions from valid (i.e., persistent) system states with each flow backlogged at at least one resource. If all flows are initially backlogged only at resource 1 and this is a persistent state, convergence has already occurred. If all flows were initially backlogged only at resource 2 we would have and . This yields a combined consumption of resource 1, , that is greater than 1 (by Lemma 6) proving that this state is in fact impossible, i.e., not persistent.

When both resources have at least one backlogged flow, the fair shares and are well-defined and characterize the system state. If flow is backlogged at 1 but not at 2, we must have , and yielding . If flow is backlogged at both resources with the queue at 1 increasing, we have , and again yielding . Similarly, for flows that are not backlogged at 1 or have a decreasing queue, we must have .

The value of the ratio partitions flows into two categories. The first denoted consists of flows such that . These flows are backlogged only at 1 or have an increasing resource 1 queue. The second category comprises the remainder and these are backlogged only at 2 or have a decreasing queue at 1. The statement of the lemma is true if flow belongs to category in every valid state.

To establish the sufficient condition, first suppose flow is not in category for some valid state since . No flow other than flows 1 to can then be in since we would require and . Let be the set of flows backlogged only at 2 and its cardinality. The fair shares and then satisfy

—

Solving we find,

The right hand side is maximized when consists of flows to for a particular value of . These are the flows with the smallest values of . To see this suppose the maximizing instead includes flow but not with ; replacing by would increase contradicting the initial assumption. Thus,

Note that and, for the positivity of and , we also require,

(9.16)

(9.17)

It is true that and would also be positive if the inequality in both (9.16) and (9.17) were inverted. We show that this is impossible for . Write

For this expression to be negative it is clearly necessary that the smallest term in the sum be negative, i.e., .

Similarly, writing

we deduce that the expression can only be negative if the smallest term in the sum is negative, i.e., if . Given the ordering of the flows following (9.5), this condition is incompatible with the previous one completing the proof that (9.16) and (9.17) are indeed necessary conditions on the value of .

Let

so that maximizes over states satisfying the above inequalities. We now show that

is actually the only possibility by successively considering all possible values of .

i) and .

If maximizes we have or

— —
(9.18)
— —

where the denominators are both positive in the assumed range. Cross-multiplying and simplifying, we find,

If, as assumed, for some valid state, then certainly implying . As and are integers, this inequality is incompatible with the assumption . We conclude from this contradiction that cannot in fact be in the considered range.

ii) — and .

We again have (9.18) but the denominator on the left is positive while the one on the right is negative. Cross-multiplying therefore inverts the inequality and we deduce,

(9.19)

Let — and — so that .
We can write

As is positive in the considered range, the inequality yields

On the other hand, is negative so that the inequality is inverted on division giving,

(9.20)

We can also write,

From (9.19) and , we have and, therefore,

This contradicts (9.20) proving, therefore, that cannot be in the considered range.

iii) .

First consider . In the state where all flows are backlogged only at resource 2, we would have and a combined flow into resource 1 of — . However, the latter expression is greater than 1 by Lemma 6 proving that this state is impossible. Flow 1 must therefore belong to .

For , in order for the initial assumption to be satisfied we require for some state. This is clearly not satisfied if . In other words, a sufficient condition for flow is indeed (9.6).

Lemma 5 - Convergence to BMF

Proof. We know from Lemma 3 that the backlog of flow 1 always converges and from Lemma 4 that the backlog of flow converges independently of if . It remains to consider states with and where the sufficient conditions for convergence are (9.6) and (9.7), respectively.

Let denote the numerator and the denominator of the right hand side of (9.6):

— —

— —

It is easily verified that and cannot both be negative for the same value of . Moreover, the values of where may be negative are the smallest possible (it is easy to see that for) and the values of where may be negative are the largest possible (since for).

The sufficient condition (9.7) that flow stabilizes to a resource 2 backlog may be expressed

If the right hand side is negative, the condition is trivially satisfied. If positive both numerator and denominator are positive and we can invert the inequality to read

If condition (9.6) is not satisfied. The proof is complete if we prove that , i.e., (9.7) is satisfied.

If then and must both be positive and . We deduce,

Now, so . Applying the same argument repeatedly we finally deduce .

Chapter 10

Conclusions

In this thesis, we have introduced, for the first time, a novel analytical approach to (in-band underlay and out-band) D2D systems based on a CPS model. A CPS is a queueing model which is able to capture the intertwined nature of user performance in a D2D-assisted cellular network. In a CPS, indeed, service rates are determined by the set of queues serving traffic. In the same way, the performance of users in a D2D-assisted cellular network is determined by the interference generated by the set of transmitters scheduled on the same resources (in-band underlay) or by the number of users competing on the MAC of WiFi (out-band).

Unfortunately, analyzing a CPS is as difficult as analyzing a D2D-assisted cellular network. We propose therefore a novel approach to CPS queuing systems, based on worst case analysis, which scales to a generic number of nodes and which enable characterizing the main performance features of (in-band underlay and out-band) D2D systems under any given operational condition, i.e., also when saturation does not hold. Interestingly, the modeling of D2D systems enables also the optimization of network operation. Leaveraging the Coupled Processor System (CPS) modeling, we propose a simple rate limiting technique which perform proportional fairness among in-band underlay and out-band D2D transmitters.

When D2D is also used for relay purposes and not only as a simple communication link among users with content to exchange, we have proposed a framework to compute throughput, airtime and power consumption of mobile terminals in a D2D-assisted cellular network. Our model is unique and accurate, and its novelty consists not only in accounting for the coupling between cellular and resource allocation/utilization in uplink and downlink with D2D relay, but also in the method used to analyze the operation by identifying network states. We have then formulated an on-line access selection problem for energy efficiency. Given the NP-hardness of the problem, we have proposed MBH, an on-line heuristic that largely outperforms existing association policies while providing enhanced levels of fairness (in terms of throughputs of terminals and energy efficiency of relay nodes). Our numerical results demonstrate that energy efficiency can be largely enhanced, and that it can be traded off for fairness while maintaining high throughput gains, thanks to the adoption of D2D-based relay.

Finally, we analyzed the performance of D2D communications when used in conjunction with other optimization techniques. We focused on handling D2D communications when ICIC techniques are in place. First, we have proposed TOMRAN, a control mechanism that combines inter-cell interference coordination with intra-cellular D2D-based traffic offloading to enhance the system spectral efficiency. We have proven that TOMRAN achieves outstanding results with respect to ICIC and D2D solutions taken in isolation, bringing substantial gain even when out-band D2D resource sharing limitations represent the communication bottleneck.

Encouraged by the TOMRAN performance results, in this manuscript we have analytically characterized the performance of a cellular network with ABSF and mm-waves D2D. In such kind of cellular network, throughput, transmission efficiency, fairness and ICIC can be addressed simultaneously by means of ABSF thanks to the adoption of ultra-high speed D2D relay techniques such as the ones based on WiGig. Our analysis shows that ABSF alone is not able to guarantee high average throughputs to all users, although it guarantees cleaner and more efficient channel transmissions. However, we have shown that ABSF can be used to tune fairness *stochastically*. We have leveraged the analysis carried out in the manuscript to define two proportional fairness optimization problems aiming to assign optimal probabilities to the occurrence of ABSF states, respectively under asymptotic and dynamic fairness targets.

In this thesis, we also argue that optimizing only the last wireless hop in 5G networks is not sufficient. Indeed, we show the importance of including in the analysis other resources which are nowadays part of the RANs: backhaul and computational resources. We consider multi-resource sharing between flows with heterogeneous resource requirements, arising in networks with wireless links or software routers implementing virtualized network functions, as C-RANs with D2D support. BMF is a sharing objective in this context that yields a satisfactory efficiency fairness tradeoff. Our aim in this thesis has been to show BMF can be realized by locally imposing fair sharing at each resource and performing end-to-end window-based flow control.

In the future, we consider the possibility of applying the theoretical results achieved in this thesis also to different scenarios than D2D networks. CPS and multi-resource sharing, indeed, are concepts which go beyond cellular networks. The proposed models can be applied, as we said, to a large number of D2D transmission modes, but also to multiprocessor systems, to problems of resource allocation in data centers and, in general, to all those systems where coupling, even distributed over different resources, is present.

References

[1] A. Osseiran, F. Boccardi, V. Braun, K. Kusume, P. Marsch, M. Maternia, O. Queseth, M. Schellmann, H. D. Schotten, T. Hidekazu, H. M. Tullberg, M. A. Uusitalo, B. Timus, and M. Fallgren, "Scenarios for 5g mobile and wireless communications: the vision of the METIS project," *IEEE Communications Magazine*, vol. 52, no. 5, pp. 26–35, 2014.

[2] M. Peng, Y. Li, Z. Zhao, and C. Wang, "System architecture and key technologies for 5g heterogeneous cloud radio access networks," *IEEE Network*, vol. 29, no. 2, pp. 6–14, 2015.

[3] G. Boudreau, J. Panicker, N. Guo, R. Chang, N. Wang, and S. Vrzic, "Interference coordination and cancellation for 4g networks," *Comm. Mag.*, vol. 47, no. 4, pp. 74–81, Apr. 2009. [Online]. Available: http://dx.doi.org/10.1109/MCOM.2009.4907410

[4] 3GPP, "Proximity-based services (prose); stage 2," 3rd Generation Partnership Project (3GPP), TS 23.303, 09 2015. [Online]. Available: http://www.3gpp.org/DynaReport/23303.htm

[5] X. Lin, J. G. Andrews, A. Ghosh, and R. Ratasuk, "An overview of 3GPP device-to-device proximity services," *IEEE Communications Magazine*, vol. 52, no. 4, pp. 40–48, Apr. 2014. [Online]. Available: http://dx.doi.org/10.1109/mcom.2014.6807945

[6] S. Xu, H. Wang, T. Chen, Q. Huang, and T. Peng, "Effective interference cancellation scheme for device-to-device communication underlaying cellular networks," in *Proceedings of IEEE VTC-Fall*, 2010, pp. 1–5.

[7] K. Doppler, M. Rinne, C. Wijting, C. B. Ribeiro, and K. Hugl, "Device-to-device communication as an underlay to lte-advanced networks," *Communications Magazine, IEEE*, vol. 47, no. 12, pp. 42–49, 2009.

[8] P. Rost, S. Talarico, and M. C. Valenti, "The complexity-rate tradeoff of centralized radio access networks," *IEEE Trans. Wireless Communications*, vol. 14, no. 11, pp. 6164–6176, 2015. [Online]. Available: http://dx.doi.org/10.1109/TWC.2015.2449321

[9] T. Bonald and J. Roberts, "Multi-resource fairness: Objectives, algorithms and performance," in *Proceedings of the 2015 ACM SIGMETRICS International Conference*

on Measurement and Modeling of Computer Systems, ser. SIGMETRICS '15. New York, NY, USA: ACM, 2015, pp. 31–42. [Online]. Available: http://doi.acm.org/10.1145/2745844.2745869

[10] G. Bianchi, "Performance analysis of the ieee 802.11 distributed coordination function," *IEEE JSAC*, vol. 18, no. 3, pp. 535–547, 2000.

[11] S. C. Borst, O. J. Boxma, and P. R. Jelenkovic, "Coupled processors with regularly varying service times," in *INFOCOM*, 2000, pp. 157–164.

[12] S. C. Borst, M. Jonckheere, and L. Leskelä, "Stability of parallel queueing systems with coupled service rates," *Discrete Event Dynamic Systems*, vol. 18, no. 4, pp. 447–472, 2008.

[13] B. Rengarajan, C. Caramanis, and G. de Veciana, "Analyzing queueing systems with coupled processors through semidefinite programming," *INFORMS: Applied Probability Session*, 2008.

[14] J.-Y. L. Boudec and P. Thiran, *Network Calculus: A Theory of Deterministic Queuing Systems for the Internet*, ser. LNCS. Springer, 2001, vol. 2050.

[15] X. Wu, S. Tavildar, S. Shakkottai, T. Richardson, J. Li, R. Laroia, and A. Jovicic, "Flashlinq: A synchronous distributed scheduler for peer-to-peer ad hoc networks," *IEEE/ACM Transactions on Networking (TON)*, vol. 21, no. 4, pp. 1215–1228, 2013.

[16] V. Paxson and S. Floyd, "Wide-area traffic: The failure of poisson modeling," *IEEE/ACM Transactions on Networking*, pp. 226–244, 1995.

[17] Nokia, "Integrated wi-fi supports superior mobile broadband," Nokia Siemens Networks, Tech. Rep., 2014. [Online]. Available: www.networks.nokia.com/system/files/document/integrated_wi-fi_supports_superior_mobile_broadband.pdf

[18] A. Asadi, Q. Wang, and V. Mancuso, "A survey on device-to-device communication in cellular networks," *IEEE Communications Surveys and Tutorials*, 2014.

[19] S. Andreev, A. Pyattaev, K. Johnsson, O. Galinina, and Y. Koucheryavy, "Cellular Traffic Offloading onto Network-Assisted Device-to-Device Connections," *IEEE Communications Magazine*, 2014.

[20] M. Cierny, H. Wang, R. Wichman, Z. Ding, and C. Wijting, "On number of almost blank subframes in heterogeneous cellular networks," *Wireless Communications, IEEE Transactions on*, vol. 12, pp. 5061–5073, October 2013.

[21] A. Asadi and V. Mancuso, "Dronee: Dual-radio opportunistic networking for energy efficiency," *Computer Communications*, vol. 50, pp. 41–52, 2014.

[22] V. Sciancalepore, V. Mancuso, A. Banchs, S. Zaks, and A. Capone, "Interference Coordination Strategies for Content Update Dissemination in LTE-A," in *The 33rd Annual IEEE International Conference on Computer Communications (INFOCOM)*, 2014.

[23] Q. Chen, J. Tang, D. Wong, X. Peng, and Y. Zhang, "Directional cooperative mac protocol design and performance analysis for ieee 802.11ad wlans," *Vehicular Technology, IEEE Transactions on*, vol. 62, no. 6, pp. 2667–2677, July 2013.

[24] P. Goyal, H. Vin, and H. Cheng, "Start-time fair queueing: a scheduling algorithm for integrated services packet switching networks," *Networking, IEEE/ACM Transactions on*, vol. 5, no. 5, pp. 690–704, Oct 1997.

[25] C. Vitale, G. Rizzo, B. Rengarajan, and V. Mancuso, "An analytical approach to performance analysis of coupled processor systems," in *27th International Teletraffic Congress, ITC 2015, Ghent, Belgium, September 8-10, 2015*, 2015, pp. 89–97. [Online]. Available: http://dx.doi.org/10.1109/ITC.2015.18

[26] C. Vitale, G. Rizzo, and V. Mancuso, "A coupled processors model for 802.11 ad hoc networks under non saturation," in *2015 IEEE International Conference on Communications, ICC 2015, London, United Kingdom, June 8-12, 2015*, 2015, pp. 628–634. [Online]. Available: http://dx.doi.org/10.1109/ICC.2015.7248392

[27] C. Vitale, V. Mancuso, and G. Rizzo, "Modelling D2D communications in cellular access networks via coupled processors," in *7th International Conference on Communication Systems and Networks, COMSNETS 2015, Bangalore, India, January 6-10, 2015*, 2015, pp. 1–8. [Online]. Available: http://dx.doi.org/10.1109/COMSNETS.2015.7098687

[28] C. Vitale and V. Mancuso, "Energy efficiency in mixed access networks," in *Proceedings of the 19th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems, MSWiM 2016, Malta, November 13-17, 2016*, 2016, pp. 35–42. [Online]. Available: http://doi.acm.org/10.1145/2988287.2989151

[29] C. Vitale, V. Sciancalepore, A. Asadi, and V. Mancuso, "Two-level opportunistic spectrum management for green 5g radio access networks," in *2015 IEEE Online Conference on Green Communications, OnlineGreenComm 2015, Piscataway, NJ, USA, November 10-12, 2015*, 2015, pp. 78–83. [Online]. Available: http://dx.doi.org/10.1109/OnlineGreenCom.2015.7387383

[30] T. Bonald, J. Roberts, and C. Vitale, "Convergence to multi-resource fairness under end-to-end window control," *The 36th Annual IEEE International Conference on Computer Communications (INFOCOM)*, 2017.

[31] C. Vitale, V. Mancuso, and G. Rizzo, "Modelling D2D communications in cellular access networks via coupled processors," in *Invited Poster: 7th International Conference*

on Communication Systems and Networks, COMSNETS 2015, Bangalore, India, January 6-10, 2015.

[32] C.-H. Yu, K. Doppler, C. B. Ribeiro, and O. Tirkkonen, "Resource sharing optimization for device-to-device communication underlaying cellular networks," *Wireless Communications, IEEE Transactions on*, vol. 10, no. 8, pp. 2752–2763, 2011.

[33] L. Lei, Y. Han, Z. Zhong, and C. Lin, "Performance analysis of device-to-device communications with frequency reuse using stochastic petri nets," in *Communications (ICC), 2013 IEEE International Conference on*. IEEE, 2013, pp. 6354–6359.

[34] P. Phunchongharn, E. Hossain, and D. I. Kim, "Resource allocation for device-to-device communications underlaying lte-advanced networks," *Wireless Communications, IEEE*, vol. 20, no. 4, 2013.

[35] M.-H. Han, B.-G. Kim, and J.-W. Lee, "Subchannel and transmission mode scheduling for D2D communication in OFDMA networks," in *Proceedings of IEEE VTC-Fall*, 2012, pp. 1–5.

[36] C. Xu, L. Song, Z. Han, Q. Zhao, X. Wang, and B. Jiao, "Interference-aware resource allocation for device-to-device communications as an underlay using sequential second price auction," in *Proceedings of IEEE ICC*, 2012, pp. 445–449.

[37] A. D. la Oliva, A. Banchs, P. Serrano, and F. A. Zdarsky, "Providing throughput guarantees in heterogeneous wireless mesh networks." *Wirel. Commun. Mob. Comput.*, 2012.

[38] F. Daneshgaran, M. Laddomada, F. Mesiti, and M. Mondin, "Unsaturated throughput analysis of ieee 802.11 in presence of non ideal transmission channel and capture effects," *Wireless Communications, IEEE Transactions on*, vol. 7, no. 4, pp. 1276–1286, April 2008.

[39] D. Malone, K. Duffy, and D. Leith, "Modeling the 802.11 distributed coordination function in nonsaturated heterogeneous conditions," *Networking, IEEE/ACM Transactions on*, vol. 15, no. 1, pp. 159–172, Feb 2007.

[40] M. Andrews, "Instability of FIFO in the permanent sessions model at arbitrarily small network loads," *ACM Transactions on Algorithms*, vol. 5, no. 3, 2009.

[41] G. Fayolle and R. Iasnogorodski, "Two coupled processors: The reduction to a riemann-hilbert problem," *Zeitschrift fur Wahrscheinlichkeitstheorie und Verwandte Gebiete*, vol. 47, no. 3, pp. 325–351, 1979. [Online]. Available: http://dx.doi.org/10.1007/BF00535168

[42] F. Guillemin and D. Pinchon, "Analysis of the weighted fair queuing system with two classes of customers with exponential service times," in *Journal of Applied Probability*, 2004.

[43] J. Pender, "An analysis of nonstationary coupled queues," *Telecommunication Systems*, vol. 61, no. 4, pp. 823–838, 2016.

[44] T. Bonald, S. C. Borst, N. Hegde, and A. Proutiére, "Wireless data performance in multi-cell scenarios," in *SIGMETRICS*, 2004, pp. 378–380.

[45] M. Jonckheere and S. C. Borst, "Stability of multi-class queueing systems with state-dependent service rates," in *VALUETOOLS*, 2006, p. 15.

[46] K. Doppler, T. Koskela, S. Hakola, and C. Ribeiro, "Enabling device-to-device communication in cellular networks," 2013, US Patent 8520575.

[47] J. Seppala, T. Koskela, T. Chen, and S. Hakola, "Network controlled device-to-device (D2D) and cluster multicast concept for LTE and LTE-A networks," in *Proc. of IEEE WCNC*, 2011.

[48] B. Zhou, H. Hu, S.-Q. Huang, and H.-H. Chen, "Intracluster device-to-device relay algorithm with optimal resource utilization," *IEEE Trans. on Vehicular Technology*, 2013.

[49] X. Bao, Y. Lin, U. Lee, I. Rimac, and R. Choudhury, "Dataspotting: Exploiting naturally clustered mobile devices to offload cellular traffic," in *Proc. of IEEE INFOCOM*, 2013.

[50] Q. Nguyen-Vuong, N. Agoulmine, and Y. Ghamri-Doudane, "A user-centric and context-aware solution to interface management and access network selection in heterogeneous wireless environments," *Computer Networks*, vol. 52, no. 18, pp. 3358 – 3372, 2008.

[51] W. Song, H. Jiang, and W. Zhuang, "Performance analysis of the wlan-first scheme in cellular/wlan interworking," *IEEE Transactions on Wireless Communications*, vol. 6, no. 5, pp. 1932–1952, 2007.

[52] S. Lee, K. Sriram, K. Kim, Y. H. Kim, and N. Golmie, "Vertical handoff decision algorithms for providing optimized performance in heterogeneous wireless networks," *IEEE T. Vehicular Technology*, vol. 58, no. 2, pp. 865–881, 2009.

[53] C. Desset, N. Ahmed, and A. Dejonghe, "Energy savings for wireless terminals through smart vertical handover," in *ICC*. IEEE, 2009, pp. 1–5.

[54] F. Malandrino, Z. Limani, C. Casetti, and C. Chiasserini, "Interference-aware downlink and uplink resource allocation in hetnets with D2D support," *IEEE Transactions on Wireless Communications*, vol. 14, no. 5, pp. 2729–2741, 2015.

[55] D. Karvounas, A. Georgakopoulos, K. Tsagkaris, V. Stavroulaki, and P. Demestichas, "Smart Management of D2D Constructs: An Experiment-Based Approach," *IEEE Communications Magazine*, 2014.

[56] A. Asadi, V. Mancuso, and R. Gupta, "An SDR-based Experimental Study of Outband D2D Communications," in *The 35th Annual IEEE International Conference on Computer Communications (INFOCOM)*, 2016.

[57] J. Qiao, X. S. Shen, J. W. Mark, Q. Shen, Y. He, and L. Lei, "Enabling device-to-device communications in millimeter-wave 5g cellular networks," *IEEE Communications Magazine*, vol. 53, no. 1, pp. 209–215, 2015. [Online]. Available: http://dx.doi.org/10.1109/MCOM.2015.7010536

[58] Third Generation Partnership Project (3GPP), "Evolved Universal Terrestrial Radio Access Network (E-UTRAN); X2 application protocol (X2AP)," 3GPP TS 36.423 v. 14.0.0, September 2016.

[59] A. Ghosh, N. Mangalvedhe, R. Ratasuk, B. Mondal, M. Cudak, E. Visotsky, T. Thomas, J. Andrews, P. Xia, H. Jo, H. S. Dhillon, and T. D. Novlan, "Heterogeneous cellular networks: From theory to practice," *IEEE Communications Magazine*, vol. 50, no. 6, pp. 54–64, 2012.

[60] A. Hamza, S. Khalifa, H. Hamza, and K. Elsayed, "A Survey on Inter-Cell Interference Coordination Techniques in OFDMA-Based Cellular Networks," *IEEE Communications Surveys Tutorials*, vol. PP, no. 99, pp. 1–29, 2013.

[61] V. Sciancalepore, V. Mancuso, and A. Banchs, "BASICS: Scheduling Base Stations to Mitigate Interferences in Cellular Networks," in *Proceedings of 14th IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, Jun. 2013.

[62] V. Sciancalepore, V. Mancuso, A. Banchs, S. Zaks, and A. Capone, "Enhanced content update dissemination through D2D in 5G cellular networks," *IEEE Transactions on Wireless Communications*, vol. 15, pp. 7517–7530, Nov 2016.

[63] M. Kamel and K. M. Elsayed, "Performance evaluation of a coordinated time-domain eICIC framework based on ABSF in heterogeneous LTE-Advanced networks," in *IEEE GLOBECOM*, 2012, pp. 5326–5331.

[64] K. Son, Y. Yi, and S. Chong, "Utility-optimal multi-pattern reuse in multi-cell networks," *IEEE Trans. on Wireless Communications*, vol. 10, no. 1, pp. 142–153, 2011.

[65] V. Sciancalepore, I. Filippini, V. Mancuso, A. Capone, and A. Banchs, "A semi-distributed mechanism for inter-cell interference coordination exploiting the absf paradigm," in *IEEE SECON*, 2015.

[66] A. Asadi, V. Sciancalepore, and V. Mancuso, "On the efficient utilization of radio resources in extremely dense wireless networks," *Communications Magazine, IEEE*, vol. 53, no. 1, pp. 126–132, January 2015.

[67] S. Singh and J. Andrews, "Joint resource partitioning and offloading in heterogeneous cellular networks," *IEEE Trans. on Wireless Communications*, vol. 13, no. 2, pp. 888–901, February 2014.

[68] Y. Niu, C. Gao, Y. Li, L. Su, D. Jin, and A. V. Vasilakos, "Exploiting device-to-device communications in joint scheduling of access and backhaul for mmwave small cells," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 10, pp. 2052–2069, 2015.

[69] E. L. Hahne and R. G. Gallager, "Round robin scheduling for fair flow control in data communication networks," in *ICC*, 1986, pp. 103–107.

[70] M. Katevenis, "Fast switching and fair control of congested flow in broadband networks," *IEEE Journal on Selected Areas in Communications*, vol. 5, no. 8, pp. 1315–1326, October 1987.

[71] E. Hahne, "Round-robin scheduling for max-min fairness in data networks," *Selected Areas in Communications, IEEE Journal on*, vol. 9, no. 7, pp. 1024–1039, Sep 1991.

[72] E. L. Hahne, "Round robin scheduling for fair flow control in data communication networks," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, USA, 1986. [Online]. Available: http://hdl.handle.net/1721.1/14932

[73] N. Chrysos and M. Katevenis, "Distributed WFQ scheduling converging to weighted maxï¿½min fairness," *Computer Networks*, vol. 55, no. 3, pp. 792–806, 2011.

[74] A. Ghodsi, M. Zaharia, B. Hindman, A. Konwinski, S. Shenker, and I. Stoica, "Dominant resource fairness: Fair allocation of multiple resource types," in *Proceedings of the 8th USENIX Conference on Networked Systems Design and Implementation*, ser. NSDI'11. Berkeley, CA, USA: USENIX Association, 2011, pp. 24–24.

[75] A. Ghodsi, V. Sekar, M. Zaharia, and I. Stoica, "Multi-resource fair queueing for packet processing," in *Proceedings of ACM SIGCOMM 2012*. New York, NY, USA: ACM, 2012, pp. 1–12.

[76] R. Srikant, *The Mathematics of Internet Congestion Control*. Birkhauser Basel, 2004.

[77] J. Nagle, "On packet switches with infinite storage," RFC 970, 1985.

[78] T. Bonald, M. Feuillet, and A. Proutiere, "Is the "law of the jungle" sustainable for the internet?" in *INFOCOM 2009, IEEE*, April 2009, pp. 28–36.

[79] P. Viswanath, D. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *Information Theory, IEEE Transactions on*, vol. 48, no. 6, pp. 1277–1294, Jun 2002.

[80] M. Heusse, F. Rousseau, G. Berger-Sabbatel, and A. Duda, "Performance anomaly of 802.11b," in *INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications. IEEE Societies*, vol. 2, March 2003, pp. 836–843 vol.2.

[81] M. Shin, S. Chong, and I. Rhee, "Dual-resource tcp/aqm for processing-constrained networks," *IEEE/ACM Trans. Netw.*, vol. 16, no. 2, pp. 435–449, Apr. 2008. [Online]. Available: http://dx.doi.org/10.1109/TNET.2007.900415

[82] Ö. B. Akan, E. Ekici, L. Qiu, and A. C. Snoeren, Eds., *The 18th Annual International Conference on Mobile Computing and Networking, Mobicom'12, Istanbul, Turkey, August 22-26, 2012*. ACM, 2012. [Online]. Available: http://dl.acm.org/citation.cfm?id=2348543

[83] M. Becchi, "From poisson processes to self-similarity: a survey of network traffic models," 2001. [Online]. Available: http://www.cse.wustl.edu/ jain/cse567-06/traffic_models1.htm

[84] Y. Bernet, S. Blake, D. Grossman, and A. Smith, "An informal management model for diffserv routers," Internet Engineering Task Force, May 2002.

[85] C.-S. Chang, *Performance Guarantees in Communication Networks*. London, UK, UK: Springer-Verlag, 2000.

[86] G. Rizzo and J. L. Boudec, "Stability and delay bounds in heterogeneous networks of aggregate schedulers," in *INFOCOM 2008. 27th IEEE International Conference on Computer Communications, Joint Conference of the IEEE Computer and Communications Societies, 13-18 April 2008, Phoenix, AZ, USA*, 2008, pp. 1490–1498. [Online]. Available: http://dx.doi.org/10.1109/INFOCOM.2008.208

[87] M. Fidler and J. B. Schmitt, "On the way to a distributed systems calculus: an end-to-end network calculus with data scaling," *SIGMETRICS*, vol. 34, no. 1, pp. 287–298, jun 2006.

[88] A. K. Parekh and R. G. Gallager, "A generalized processor sharing approach to flow control in integrated services networks: the single-node case," *IEEE/ACM Trans. Netw.*, vol. 1, no. 3, pp. 344–357, Jun. 1993.

[89] S. Borst, N. Hegde, and A. Proutiere, "Interacting queues with server selection and coordinated scheduling: application to cellular data networks," *Annals of Operations Research*, vol. 170, no. 1, pp. 59–78, 2009.

[90] D. Wang and X. Wang, "Effective interference cancellation schemes for device-to-device multicast uplink period underlaying cellular networks," *Wireless Personal Communications*, vol. 75, no. 4, pp. 2201–2216, 2014.

[91] P. Jänis, C. Yu, K. Doppler, C. B. Ribeiro, C. Wijting, K. Hugl, O. Tirkkonen, and V. Koivunen, "Device-to-device communication underlaying cellular communications systems," *IJCNS*, vol. 2, no. 3, pp. 169–178, 2009.

[92] X. Lin, J. G. Andrews, A. Ghosh, and R. Ratasuk, "An overview on 3gpp device-to-device proximity services," *CoRR*, vol. abs/1310.0116, 2013. [Online]. Available: http://arxiv.org/abs/1310.0116

[93] J. S. Seybold, *Introduction to RF propagation*. J. Wiley & Sons, 2005.

[94] O. Iosif and I. Banica, "Lte uplink analysis using two packet scheduling models," in *Telecommunications Forum (TELFOR), 2011 19th*. IEEE, 2011, pp. 394–397.

[95] A. Vecchietti, S. Lee, and I. E. Grossmann, "Modeling of discrete/continuous optimization problems: characterization and formulation of disjunctions and their relaxations," *Computers and Chemical Engineering*, vol. 27, no. 3, pp. 433–448, 2003.

[96] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, USA: Cambridge University Press, 2004.

[97] V. Gazis, "A survey of standards for machine-to-machine and the internet of things," *IEEE Communications Surveys Tutorials*, vol. 19, no. 1, pp. 482–511, 2017.

[98] D. Neves da Hora, K. Van Doorselaer, K. Van Oost, R. Teixeira, and C. Diot, "Passive wi-fi link capacity estimation on commodity access points," in *Traffic Monitoring and Analysis Workshop (TMA) 2016*, 2016, pp. 394–397.

[99] I. L. S. Committee *et al.*, "Ieee 802.11-wireless lan medium access control (mac) and physical layer (phy) specifications," in *IEEE, June*, 2007.

[100] 3GPP, "Evolved universal terrestrial radio access (e-utra); physical layer procedures (release 12)," 3rd Generation Partnership Project (3GPP), TS 36.213, 2015. [Online]. Available: http://www.3gpp.org/dynareport/36213.htm

[101] T. Israeli, E. Biton, and O. Gurewitz, "Experimental assessment of power-save behavior of commercial ieee 802.16 network," *IEEE Transactions on Wireless Communications*, vol. 14, no. 3, pp. 1169–1182, March 2015.

[102] B. Dusza, C. Ide, L. Cheng, and C. Wietfeld, "An accurate measurement-based power consumption model for LTE uplink transmissions," in *IEEE INFOCOM (Poster)*. Turin, Italy: IEEE, April 2013.

[103] A. Garcia-Saavedra, P. Serrano, A. Banchs, and G. Bianchi, "Energy consumption anatomy of 802.11 devices and its implication on modeling and design," in *CoNEXT*, 2012, pp. 169–180.

[104] D. G. Cattrysse and L. N. Van Wassenhove, "A survey of algorithms for the generalized assignment problem," *European Journal of Operational Research*, vol. 60, no. 3, pp. 260–272, 1992.

[105] K. Zeineddine, M. L. Honig, and S. Nagaraj, "Uplink power allocation for distributed interference cancellation with channel estimation error," *IEEE Transactions on Wireless Communications*, vol. 15, no. 10, pp. 6785–6796, 2016.

[106] D. Korpi, L. Anttila, V. Syrjala, and M. Valkama, "Widely linear digital self-interference cancellation in direct-conversion full-duplex transceiver," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 9, pp. 1674–1687, 2014.

[107] M. Sauter, *From GSM to LTE-advanced: An Introduction to Mobile Networks and Mobile Broadband*. John Wiley & Sons, 2014.

[108] Wi-Fi Alliance Specification, "Wi-Fi Peer-to-Peer (P2P) Specification v1.1," 2011.

[109] ITU, "Guidelines for evaluation of radio interface technologies for IMT-Advanced," 2009.

[110] X. Hong, M. Gerla, G. Pei, and C.-C. Chiang, "A group mobility model for ad hoc wireless networks," in *Proceedings of the 2nd ACM International Workshop on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, 1999, pp. 53–60.

[111] Third Generation Partnership Project (3GPP), "Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Radio Frequency (RF) system scenarios," 3GPP TS 36.942 v. 13.0.0, January 2016.

[112] D. P. Meyer and H. A. Mayer, "Radar target detection- handbook of theory and practice," *New York, Academic Press, Inc., 1973. 508 p*, 1973.

[113] C. Bettstetter, G. Resta, and P. Santi, "The node distribution of the random waypoint mobility model for wireless ad hoc networks," *Mobile Computing, IEEE Transactions on*, vol. 2, no. 3, pp. 257–269, July 2003.

[114] P. Rode, C. Hoffman, J. Kandt, D. Smith, and A. Graff, "Towards new urban mobility the case of london and berlin," *LSE Cities*, Sep 2015.

[115] B. Radunovic and J. Y. L. Boudec, "A unified framework for max-min and min-max fairness with applications," *IEEE/ACM Transactions on Networking*, vol. 15, no. 5, pp. 1073–1083, Oct 2007.

[116] "802.11-2012 - IEEE Standard for Information technology–Telecommunications and information exchange between systems Local and metropolitan area networks–Specific requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications," IEEE-Inst, Tech. Rep. IEEE Std 802.11-2012. [Online]. Available: http://ieeexplore.ieee.org/servlet/opac?punumber=6178209